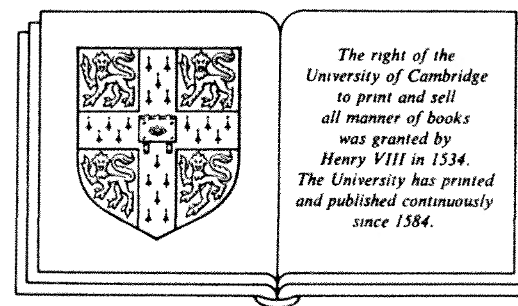
Matrix analysis

ROGER A. HORN
The Johns Hopkins University
CHARLES R. JOHNSON
College of William and Mary



CAMBRIDGE UNIVERSITY PRESS
Cambridge
London New York New Rochelle
Melbourne Sydney

To the matrix theory community
and
to our families

Dana, Jennifer, and Emily
Susan, Ceres, Corinne, and Howard

for their understanding support

Published by the Press Syndicate of the University of Cambridge
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
40 West 20th Street, New York, NY 10011, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1985
Reprinted with corrections 1987
Reprinted 1988
Reprinted with corrections 1990

Printed in the United States of America

Library of Congress Cataloging in Publication Data

Horn, Roger A.

Matrix analysis.

Bibliography: p.
Includes index.

1. Matrices. I. Johnson, Charles R. II. Title
QA188.H66 1985 512.9'434 85-7736
ISBN 0-521-30586-1 hard covers
ISBN 0-521-38632-2 paperback

Contents

Preface

Linear algebra and matrix theory have long been fundamental tools in mathematical disciplines as well as fertile fields for research in their own right. In this book, and in the companion volume, *Topics in Matrix Analysis,* we present classical and recent results of matrix analysis that have proved to be important to applied mathematics. The book may be used as an undergraduate or graduate text and as a self-contained reference for a variety of audiences. We assume background equivalent to a one-semester elementary linear algebra course and knowledge of rudimentary analytical concepts. We begin with the notions of eigenvalues and eigenvectors; no prior knowledge of these concepts is assumed.

Facts about matrices, beyond those found in an elementary linear algebra course, are necessary to understand virtually any area of mathematical science, whether it be differential equations; probability and statistics; optimization; or applications in theoretical and applied economics, the engineering disciplines, or operations research, to name only a few. But until recently, much of the necessary material has occurred sporadically (or not at all) in the undergraduate and graduate curricula. As interest in applied mathematics has grown and more courses have been devoted to advanced matrix theory, the need for a text offering a broad selection of topics has become more apparent, as has the need for a modern reference on the subject.

There are a number of well-loved classics in matrix theory, but they are not well suited for general classroom use, nor for systematic individual study. A lack of problems, applications, and motivation; an inadequate index; and a dated approach are among the difficulties confronting readers of some traditional references. More recent books tend to be either

elementary texts or treatises devoted to special topics. Our goal was to write a book that would be a useful modern treatment of a broad range of topics.

One view of "matrix analysis" is that it consists of those *topics* in linear algebra that have arisen out of the needs of mathematical analysis, such as multivariable calculus, complex variables, differential equations, optimization, and approximation theory. Another view is that matrix analysis is an *approach* to real and complex linear algebraic problems that does not hesitate to use notions from analysis - such as limits, continuity, and power series - when these seem more efficient or natural than a purely algebraic approach. Both views of matrix analysis are reflected in the choice and treatment of topics in this book. We prefer the term *matrix analysis* to *linear algebra* as an accurate reflection of the broad scope and methodology of the field.

For review and convenience in reference, Chapter 0 contains a summary of necessary facts from elementary linear algebra, as well as other useful, though not necessarily elementary, facts. Chapters 1, 2, and 3 contain mainly core material likely to be included in any second course in linear algebra or matrix theory: a basic treatment of eigenvalues, eigenvectors, and similarity; unitary similarity, Schur triangularization and its implications, and normal matrices; and canonical forms and factorizations including the Jordan form, *LU* factorization, *QR* factorization, and companion matrices. Beyond this, each chapter is developed substantially independently and treats in some depth a major topic:

> *Hermitian and complex symmetric matrices* (Chapter 4). We give special emphasis to variational methods for studying eigenvalues of Hermitian matrices and include an introduction to the notion of majorization.
>
> *Norms on vectors and matrices* (Chapter 5) are essential for error analyses of numerical linear algebraic algorithms and for the study of matrix power series and iterative processes. We discuss the algebraic, geometric, and analytic properties of norms in some detail, and make a careful distinction between those norm results for matrices that depend on the submultiplicativity axiom for matrix norms and those that do not.
>
> *Eigenvalue location and perturbation results* (Chapter 6) for general (not necessarily Hermitian) matrices are important for many applications. We give a detailed treatment of the theory of Geršgorin regions, and some of its modern refinements, and of relevant graph theoretic concepts.

> *Positive definite matrices* (Chapter 7) and their applications, including inequalities, are considered at some length. A discussion of the polar and singular value decompositions is included, along with applications to matrix approximation problems.
>
> *Component-wise nonnegative and positive matrices* (Chapter 8) arise in many applications in which nonnegative quantities necessarily occur (probability, economics, engineering, etc.), and their remarkable theory reflects the applications. Our development of the theory of nonnegative, positive, primitive, and irreducible matrices proceeds in elementary steps based upon the use of norms.

In the companion volume, further topics of similar interest are treated: the field of values and generalizations; inertia, stable matrices, *M*-matrices and related special classes; matrix equations, Kronecker and Hadamard products; and various ways in which functions and matrices may be linked.

This book provides the basis for a variety of one- or two-semester courses through selection of chapters and sections appropriate to a particular audience. We recommend that an instructor make a careful preselection of sections and portions of sections of the book for the needs of a particular course. This would probably include Chapter 1, much of Chapters 2 and 3, and facts about Hermitian matrices and norms from Chapters 4 and 5.

Most chapters contain some relatively specialized or nontraditional material. For example, Chapter 2 includes not only Schur's basic theorem on unitary triangularization of a single matrix, but also a discussion of simultaneous triangularization of families of matrices. In the section on unitary equivalence, our presentation of the usual facts is followed by a discussion of trace conditions for two matrices to be unitarily equivalent. A discussion of complex symmetric matrices in Chapter 4 provides a counterpoint to the development of the classical theory of Hermitian matrices. Basic aspects of a topic appear in the initial sections of each chapter, while more elaborate discussions occur at the ends of sections or in later sections. This strategy has the advantage of presenting topics in a sequence that enhances the book's utility as a reference. It also provides a rich variety of options to the instructor.

Many of the results discussed hold or can be generalized to hold for matrices over other fields or in some broader algebraic setting. However, we deliberately confine our domain to the real and complex fields where familiar methods of classical analysis as well as formal algebraic techniques may be employed.

Though we generally consider matrices to have complex entries, most examples are confined to real matrices, and no deep knowledge of complex analysis is required. Acquaintance with the arithmetic of complex numbers is necessary for an understanding of matrix analysis and is covered to the extent necessary in an appendix. Other brief appendices cover several peripheral, but essential, topics such as Weierstrass's theorem and convexity.

We have included many exercises and problems because we feel these are essential to the development of an understanding of the subject and its implications. The exercises occur throughout as part of the development of each section; they are generally elementary and of immediate use in understanding the concepts. We recommend that the reader work at least a broad selection of these. Problems are listed (in no particular order) at the end of each section; they cover a range of difficulties and types (from theoretical to computational) and they may extend the topic, develop special aspects, or suggest alternate proofs of major ideas. Significant hints are given for the more difficult problems. The results of some problems are referred to in other problems or in the text itself. We cannot overemphasize the importance of the reader's active involvement in carrying out the exercises and solving problems.

While the book itself is not about applications, we have, for motivational purposes, begun each chapter with a section outlining a few applications to introduce the topic of the chapter.

Readers who wish to consult alternate treatments of a topic for additional information are referred to the books listed in the References section following the appendices. These books are cited in the text using a brief mnemonic code; for example, a book by Jones and Smith might be referred to as [JSm]. The codes and complete citations appear alphabetically by author in the References section.

The list of book references is not exhaustive. As a practical concession to the limits of space in a general multitopic book, we have minimized the number of citations in the text. A small selection of references to papers – such as those we have explicitly used – does occur at the end of most sections accompanied by a brief discussion, but we have made no attempt to collect historical references to classical results. Extensive bibliographies are provided in the more specialized books we have referenced. The reader should also be aware of broad and current bibliographical resources covering portions of matrix analysis such as the *KWIC Index for Numerical Linear Algebra* [CaLe] and sections 15 and 65 of the *Mathematical Reviews.*

We appreciate the helpful suggestions of our colleagues and students who have taken the time to convey their reactions to the class notes and

preliminary manuscripts that were the precursors of the book. They include Wayne Barrett, Leroy Beasley, Bryan Cain, David Carlson, Dipa Choudhury, Risana Chowdhury, Yoo Pyo Hong, Dmitry Krass, Dale Olesky, Stephen Pierce, Leiba Rodman, and Pauline van den Driessche.

R.A.H.
C.R.J.

CHAPTER 0

Review and miscellanea

0.0 Introduction

The purpose of this chapter is to catalog briefly, without proof, a number of useful concepts and facts, many of which implicitly or explicitly underlie the material covered in the main portion of the book. Much of this material would be included, in some form, in an elementary course in linear algebra, but we also include a number of useful items that are not commonly found elsewhere or that do not easily fit into the subsequent structure. Thus, this section may serve the reader as a short review prior to beginning the book or as a convenient reference when necessary. We also use this chapter to set basic notation and give some definitions; thus, reference to it will also be useful for these purposes. We do assume that the reader is already familiar with the elementary concepts of linear algebra and with mechanical aspects of matrix manipulations, such as matrix multiplication and addition.

0.1 Vector spaces

Though generally implicitly, and not usually explicitly, involved in the treatment in this book, a vector space is the fundamental setting for matrix theory.

0.1.1 **Scalar field.** Underlying a vector space is the *field,* or set of scalars, from which multiplication occurs. For our purposes, that underlying field will almost always be the real numbers **R** or the complex numbers **C** (see Appendix A) under the usual addition and multiplication, but

it could be the rational numbers, the integers modulo a specified prime number, or some other field. When the field is unspecified, we use the symbol **F**. To qualify as a field, a set of scalars must be closed under two specified binary operations ("addition" and "multiplication"); both operations must be associative and commutative and have an identity element in the set; inverses must exist in the set for all elements under the addition operation and for all elements except the additive identity (0) under the multiplication operation; the multiplication operation must also be distributive over the addition operation.

0.1.2 **Vector spaces.** A *vector space* V over a field **F** is a set V of objects (called vectors) which is closed under a binary operation ("addition") which is associative and commutative and has an identity ("0") and additive inverses in the set. The set is also closed under an operation of left multiplication of the vectors by elements of the scalar field **F**, with the following properties for all $a, b \in \mathbf{F}$ and all $x, y \in V$: $a(x+y) = ax+ay$, $(a+b)x = ax+bx$, $a(bx) = (ab)x$, and $ex = x$ for the multiplicative identity $e \in \mathbf{F}$.

For a given field **F**, the set $\mathbf{F}^n$ of n-tuples (n a positive integer) with components from **F** forms a vector space over **F** under the obvious operations (component-wise addition in $\mathbf{F}^n$). The special cases $\mathbf{R}^n$ and $\mathbf{C}^n$ are the basic vector spaces of this book. The set of polynomials with real or with complex coefficients (of no more than a specified degree *or* of arbitrary degree) and the set of real or complex valued continuous functions or arbitrary functions on an interval $[a,b] \subset \mathbf{R}$ are also examples of vector spaces (over **R** or **C**). There is, of course, a fundamental difference between the finite-dimensional space $\mathbf{R}^n$ and the infinite-dimensional vector space of real-valued continuous functions on $[0,1]$.

0.1.3 **Subspaces and span.** A *subspace* U of a vector space V is a subset of V that is, by itself, a vector space over the same scalar field. For example, $\{[a,b,0]^T : a, b \in \mathbf{R}\}$ is a subspace of $\mathbf{R}^3$. Usually a subspace of a vector space V is defined by some relation that identifies particular elements of V in such a way that the resulting set is closed under the addition in V – for example, the elements of $\mathbf{R}^3$ with last component 0. It is in this regard that it is useful to think of the resulting set as a subspace rather than as a vector space in its own right. In any event, the intersection of two subspaces is again a subspace.

If S is a subset of a vector space V, the *span* of S is the set Span $S \equiv \{a_1 v_1 + a_2 v_2 + \cdots + a_k v_k : a_1, \ldots, a_k \in \mathbf{F}, v_1, \ldots, v_k \in S, k = 1, 2, \ldots\}$. Notice that Span S is always a subspace even if S is not a subspace. The set S is said to span the vector space V if Span $S = V$.

0.1.4 **Linear dependence and independence.** A set of vectors $\{x_1, x_2, \ldots, x_k\}$ in a vector space is said to be *linearly dependent* if there exist coefficients $a_1, \ldots, a_k$, not all 0, in the underlying scalar field **F** such that

$$a_1 x_1 + a_2 x_2 + \cdots + a_k x_k = 0$$

Equivalently, one of the x_i terms is a linear combination, with coefficients from **F**, of the others. For example $\{[1,2,3]^T, [1,0,-1]^T, [2,2,2]^T\}$ is a linearly dependent set in $\mathbf{R}^3$. A subset of V that is not linearly dependent over **F** is said to be *linearly independent*. For example, $\{[1,2,3]^T, [1,0,-1]^T\}$ is a linearly independent set in $\mathbf{R}^3$. It is important to note that both concepts intrinsically pertain to *sets* of vectors. Any subset of a linearly independent set is linearly independent; $\{0\}$ is a linearly dependent set; and hence any set which includes the 0 vector is linearly dependent. It can happen that a set of vectors is linearly dependent, while any proper subset of it is linearly independent.

0.1.5 **Basis.** A subset S of a vector space V is said to *span* V if every element of V may be represented as a linear combination (with coefficients from the underlying scalar field) of elements of S. For example, $\{[1,0,0]^T, [0,1,0]^T, [0,0,1]^T, [1,0,-1]^T\}$ spans $\mathbf{R}^3$ over **R** (or $\mathbf{C}^3$ over **C**). A linearly independent set which spans a vector space V is called a *basis* for V. Bases are highly nonunique, but are very efficient in that each element of V can be represented in terms of the basis in one and only one way, and this is no longer true if any element whatsoever is appended to or deleted from the basis. An independent set in V is a basis of V if and only if no set which properly contains it is independent. A set that spans V is a basis for V if and only if no proper subset of it still spans V. Every vector space has a basis.

0.1.6 **Extension to a basis.** Any linearly independent set in a vector space V may be extended to a basis of V; that is, given a linearly independent set $\{x_1, x_2, \ldots, x_k\}$ in V, there exist additional vectors $x_{k+1}, \ldots, x_n, \ldots \in V$ such that $\{x_1, \ldots, x_n, \ldots\}$ is a basis of V. The extension of a given independent set to a basis is, of course, not unique [for example, any vector with nonzero third component may be appended to the independent set $\{[1,0,0]^T, [0,1,0]^T\}$ to produce a basis of $\mathbf{R}^3$]. The example of the real vector space $C[0,1]$ of real-valued continuous functions on $[0,1]$ shows that a basis need not, in general, be finite; the infinite set of monomials $\{1, x, x^2, x^3, \ldots\}$ is an independent set in $C[0,1]$.

0.1.7 **Dimension.** If some basis of the vector space V consists of a finite number of elements, then all bases have the same number of

elements; this common number is called the *dimension* of the vector space V, and is denoted by dim V. In this event, V is said to be finite-dimensional; otherwise V is said to be infinite-dimensional. In the infinite-dimensional case (e.g., $C[0, 1]$), there is a one-to-one correspondence between the elements of any two bases. The real vector space $\mathbf{R}^n$ has dimension n. The vector space $\mathbf{C}^n$ has dimension n over the field $\mathbf{C}$ but has dimension $2n$ over the field $\mathbf{R}$. The basis $\{e_1, e_2, \ldots, e_n\}$ in which e_i has a 1 as its ith component and 0's elsewhere is sometimes called the *standard basis* of $\mathbf{R}^n$ or $\mathbf{C}^n$.

0.1.8 **Isomorphism.** If U and V are vector spaces over the same scalar field $\mathbf{F}$, and if $f: U \to V$ is an invertible function such that $f(ax+by) = af(x)+bf(y)$ for all $x, y \in U$ and all $a, b \in \mathbf{F}$, then f is said to be an *isomorphism* and U and V are said to be isomorphic ("same-structure"). Two finite-dimensional vector spaces over the same field are isomorphic if and only if they have the same dimension; thus, any n-dimensional vector space over the field $\mathbf{F}$ is isomorphic to $\mathbf{F}^n$. Any n-dimensional real vector space is, therefore, isomorphic to $\mathbf{R}^n$, and any n-dimensional complex vector space is isomorphic to $\mathbf{C}^n$. Specifically, if V is an n-dimensional vector space over a field $\mathbf{F}$ with specified basis $\mathcal{B} = \{x_1, \ldots, x_n\}$, then, since any element $x \in V$ may be written uniquely as $x = a_1x_1 + \cdots + a_nx_n$, $a_i \in \mathbf{F}$, $i = 1, \ldots, n$, we may associate x with the n-tuple $[x]_{\mathcal{B}} = [a_1, \ldots, a_n]^T$, relative to the basis $\mathcal{B}$. The mapping $x \to [x]_{\mathcal{B}}$ is an isomorphism between V and $\mathbf{F}^n$ for any basis $\mathcal{B}$.

0.2 Matrices

The fundamental object of study here may be thought of in two important ways: as a rectangular array of scalars and as a linear transformation between two vector spaces, given specified bases for each space.

0.2.1 **Rectangular arrays.** A *matrix* is an m-by-n array of scalars from a field $\mathbf{F}$. If $m = n$, the matrix is said to be square. The set of all m-by-n matrices over $\mathbf{F}$ is denoted by $M_{m,n}(\mathbf{F})$, and $M_{n,n}(\mathbf{F})$ is abbreviated to $M_n(\mathbf{F})$. In the most common case in which $\mathbf{F} = \mathbf{C}$, the complex numbers, $M_n(\mathbf{C})$ is further abbreviated to M_n, and $M_{m,n}(\mathbf{C})$ to $M_{m,n}$. Matrices are usually denoted by capital letters. For example, if

$$A = \begin{bmatrix} 2 & -\frac{3}{2} & 0 \\ -1 & \pi & 4 \end{bmatrix}$$

then $A \in M_{2,3}(\mathbf{R})$. A *submatrix* of a given matrix is a rectangular array lying in specified subsets of the rows and columns of a given matrix.

For example $[\pi \;\; 4]$ is a submatrix (lying in row 2 and columns 2 and 3) of A, above.

0.2.2 **Linear transformations.** Let U be an n-dimensional vector space and V be an m-dimensional vector space over the same scalar field $\mathbf{F}$; let $\mathcal{B}_U$ be a basis of U and $\mathcal{B}_V$ be a basis of V. We may use the isomorphisms $x \to [x]_{\mathcal{B}_U}$ and $y \to [y]_{\mathcal{B}_V}$ to represent vectors in U and V as n-tuples and m-tuples over $\mathbf{F}$, respectively. A linear transformation is a function $T: U \to V$ such that $T(a_1x_1 + a_2x_2) = a_1T(x_1) + a_2T(x_2)$ for arbitrary scalars a_1 and a_2 and vectors x_1 and x_2. A matrix $A \in M_{m,n}(\mathbf{F})$ corresponds to a linear transformation $T: U \to V$ in the following way: The vector $y = T(x)$ if and only if $[y]_{\mathcal{B}_V} = A[x]_{\mathcal{B}_U}$. The matrix A is said to represent the linear transformation T (relative to the bases $\mathcal{B}_U$ and $\mathcal{B}_V$); the representing matrix A depends upon the bases chosen. When we study the matrix A, we realize we are studying a linear transformation relative to a particular choice of bases, but explicit appeal to the bases is usually not necessary.

0.2.3 **Vector spaces associated with a given matrix or linear transformation.** There is no loss of generality in associating an n-dimensional vector space over $\mathbf{F}$ with $\mathbf{F}^n$, and we shall think of $A \in M_{m,n}(\mathbf{F})$ as a linear transformation from $\mathbf{F}^n$ to $\mathbf{F}^m$ (and also as an array). The *domain* of such a linear transformation is $\mathbf{F}^n$; its *range* is $\{y \in \mathbf{F}^m : y = Ax$ for some $x \in \mathbf{F}^n\}$. The *null space* of A is $\{x \in \mathbf{F}^n : Ax = 0\}$. The range of A is a subspace of $\mathbf{F}^m$, and the null space of A is a subspace of $\mathbf{F}^n$. We have the relation

$$n = \text{dimension of null space of } A + \text{dimension of the range of } A$$

between these two subspaces.

0.2.4 **Matrix operations.** Matrix addition is defined entry-wise for arrays of the same dimensions and is denoted by + ("$A+B$"). It corresponds to addition of linear transformations (relative to the same basis), and it inherits commutativity and associativity from the scalar field. The zero matrix (all entries zero) is the identity under addition, and $M_{m,n}(\mathbf{F})$ is itself a vector space over $\mathbf{F}$. Matrix multiplication is defined in the usual way, is denoted by juxtaposition, AB, and corresponds to the composition of linear transformations. As such, it is defined only when $A \in M_{m,n}(\mathbf{F})$, $B \in M_{p,q}(\mathbf{F})$, and $p = n$; it is associative. It is not, in general, commutative, for example,

$$\begin{bmatrix}1&0\\0&2\end{bmatrix}\begin{bmatrix}1&2\\3&4\end{bmatrix}\neq\begin{bmatrix}1&2\\3&4\end{bmatrix}\begin{bmatrix}1&0\\0&2\end{bmatrix}$$

but it can be commutative when restricted to certain subsets of $M_n(\mathbf{F})$, which are worthy of study. There is an identity under matrix multiplication, the matrix $I \in M_n(\mathbf{F})$ of the form

$$I=\begin{bmatrix}1 & & & 0\\ & 1 & & \\ & & \ddots & \\ 0 & & & 1\end{bmatrix}$$

This matrix and all scalar multiples of it (called scalar matrices) commute with all other matrices in $M_n(\mathbf{F})$ and are the only matrices which do so. Matrix multiplication is distributive over matrix addition.

We note here that the symbol 0 is used throughout to denote each of the following: the zero scalar, the zero vector (all components equal to the zero scalar), and the zero matrix (all entries equal to the zero scalar). Generally, the context will make clear which it is, so that confusion need not result. We also use the symbol I to denote the identity matrix of any size. If there is potential for confusion, the dimension will be indicated.

0.2.5 **The transpose and Hermitian adjoint.** If $A=[a_{ij}]\in M_{m,n}(\mathbf{F})$, the *transpose* of A, denoted A^T, is that matrix in $M_{n,m}(\mathbf{F})$ whose entries are a_{ji}; that is, rows are exchanged for columns and vice versa. For example,

$$\begin{bmatrix}1&2&3\\4&5&6\end{bmatrix}^T=\begin{bmatrix}1&4\\2&5\\3&6\end{bmatrix}$$

Of course, $(A^T)^T=A$. The *Hermitian adjoint* A^* of $A\in M_{m,n}(\mathbf{C})$ is defined by $A^*=\bar{A}^T$, where $\bar{A}$ is the component-wise conjugate. For example,

$$\begin{bmatrix}1+i&2-i\\-3&-2i\end{bmatrix}^*=\begin{bmatrix}1-i&-3\\2+i&2i\end{bmatrix}$$

Both the transpose and the Hermitian adjoint [and the inverse to be discussed in (0.5)] obey the *reverse-order law:* $[AB]^*=B^*A^*$ and $(AB)^T=B^TA^T$, assuming the product is defined. For the conjugate of a product, there is no reversing: $\overline{AB}=\bar{A}\bar{B}$. If $x, y\in M_{n,1}=\mathbf{C}^n$, then y^*x is a scalar, and its Hermitian adjoint and complex conjugate are the same; thus, $(y^*x)^*=\overline{y^*x}=x^*y=y^T\bar{x}$.

0.2.6 **Metamechanics of matrix multiplication.** We mention here some simple features of matrix multiplication that are useful again and again.

1. If b_j denotes the jth column of the matrix B, then the jth column of the product AB is just Ab_j.
2. If a_i denotes the ith row of the matrix A, then the ith row of the product AB is just a_iB.

To paraphrase, in the product AB, left multiplication by A multiplies the columns of B and right multiplication by B multiplies the rows of A. Further observations of this type, when one of the factors is diagonal, are made in (0.9.1).

3. If $A\in M_{m,n}(\mathbf{F})$ and $x\in\mathbf{F}^n$, then Ax is a linear combination of the columns of A (the coordinates of x are the coefficients).
4. If $A\in M_{m,n}(\mathbf{F})$ and $y\in\mathbf{F}^m$, then y^TA is a linear combination of the rows of A (the coordinates of y are the coefficients).

0.3 Determinants

Often in mathematics it is useful to summarize a multivariate phenomenon with a single number, and the determinant is an example of this. It is defined only for square matrices $A\in M_n(\mathbf{F})$, and it may be presented in two important, apparently different, but equivalent ways. We denote the determinant of $A\in M_n(\mathbf{F})$ by $\det A$.

0.3.1 **Laplace expansion.** The determinant may be defined inductively for $A=[a_{ij}]\in M_n(\mathbf{F})$ in the following way. Assume the determinant is defined over $M_{n-1}(\mathbf{F})$ and let $A_{ij}\in M_{n-1}(\mathbf{F})$ denote the submatrix of $A\in M_n(\mathbf{F})$ resulting from the deletion of row i and column j. Then

$$\sum_{j=1}^{n}(-1)^{i+j}a_{ij}\det A_{ij}=\sum_{i=1}^{n}(-1)^{i+j}a_{ij}\det A_{ij}$$

for all $i\le n$, $j\le n$, and this common value is $\det A$. The left-hand side is the Laplace expansion by minors along row i, and the right-hand side is the Laplace expansion along column j [see (0.7.1)]. For any choice of row or column, either expansion yields the determinant. This inductive presentation begins by defining the determinant of a 1-by-1 matrix to be the value of the single entry. Thus,

$$\det[a_{11}] = a_{11}$$

$$\det\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

$$\det\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31}$$

and so on. It is also clear that $\det A^T = \det A$ and that $\det A^* = \overline{\det A}$ if $A \in M_n(\mathbf{C})$.

0.3.2 **Alternating sum.** As motivated by the low-dimensional examples above, we also have for $A = [a_{ij}] \in M_n(\mathbf{F})$ that

$$\det A = \sum_{\sigma} \operatorname{sgn} \sigma \prod_{i=1}^{n} a_{i\sigma(i)}$$

where the sum runs over all $n!$ permutations σ of the n items $\{1, \ldots, n\}$ and the "sign" or "signum" of a permutation σ, $\operatorname{sgn} \sigma$, is $+1$ or -1, according to whether the minimum number of transpositions, or pair-wise interchanges, necessary to achieve it starting from $\{1, 2, \ldots, n\}$ is even or odd. Thus, each product

$$a_{1\sigma(1)} a_{2\sigma(2)} \cdots a_{n\sigma(n)}$$

enters into the determinant with a $+$ sign if the permutation σ is even or a $-$ sign if it is odd.

If the coefficient $\operatorname{sgn} \sigma$ is replaced by certain other functions, the so-called *generalized matrix functions* result in place of $\det A$. One example is per A, the *permanent* of A, in which $\operatorname{sgn} \sigma$ is replaced by the function which is identically 1.

0.3.3 **Elementary operations.** Three simple and fundamental operations may be used to put any matrix in a simple, unique (i.e., canonical) form appropriate to it for purposes such as solving linear equations, calculating determinants, matrix inversion, and determination of rank. Concentrating upon rows, they are as follows.

Type 1: Interchange of two rows

The interchange of rows i and j may be effected via left multiplication by the matrix

$$\begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & 0 & & & 1 & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & 1 & & & & 0 & \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{bmatrix}$$

(row i contains the 0 in column i and the 1 in column j; row j contains the 1 in column i and the 0 in column j)

column i column j — row i, row j

in which the two off-diagonal 1's are in the i, j and j, i positions and all unspecified entries are 0.

Type 2: Multiplication of a row by a nonzero scalar

Multiplication of row i of A by the scalar c may be accomplished via left multiplication by the matrix

$$\begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & c & & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{bmatrix}$$

(row i, column i)

in which the scalar c occurs in the i, i position.

Type 3: Addition of a scalar multiple of one row to another row

Addition of c times row i to row j results from left multiplication of A by the matrix

$$\begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & c & & 1 \end{bmatrix} \quad \text{row } j$$

column i

in which the scalar c occurs in the j, i position. Note that the matrices of each of the three elementary operations are just the result of the operation applied to the identity matrix I.

The effect of a type 1 operation upon the determinant is to multiply it by -1; the effect of a type 2 operation is to multiply it by the scalar c; a type 3 elementary operation does not change the determinant. It follows that a matrix with a zero row, or with two dependent rows, or with any k rows dependent, has determinant zero. A matrix has determinant zero if and only if a subset of its rows is linearly dependent.

0.3.4 **Row-reduced echelon form.** To each $A \in M_{m,n}(\mathbf{F})$ there corresponds a canonical form in $M_{m,n}(\mathbf{F})$, the *row-reduced echelon form* (RREF) of A, which may be attained by a (nonunique) sequence of elementary operations. Many matrices have the same RREF, but each matrix has only one RREF regardless of the sequence of elementary operations used to attain it. The defining specifications of the RREF are:

(a) Each nonzero row has 1 as its first nonzero entry;
(b) All other entries in the column of such a leading 1 are equal to 0;
(c) Any rows consisting entirely of zeroes occur at the bottom of the matrix; and
(d) The leading 1's occur in a "stairstep" pattern, left to right; that is, a leading 1 in a lower row must occur to the right of its counterpart above it.

For example,

$$\begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 & 0 & \pi \\ 0 & 0 & 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

is in RREF. The determinant of $A \in M_n(\mathbf{F})$ is nonzero if and only if its RREF is the identity

$$I = \begin{bmatrix} 1 & & 0 \\ & 1 & \\ & & \ddots & \\ 0 & & & 1 \end{bmatrix}$$

(whose determinant is 1). The value of $\det A$ may be calculated by recording the effects upon the determinant of each of the elementary operations which lead to the RREF.

For the system of linear equations $Ax = b$ with $A \in M_{m,n}(\mathbf{F})$ and $b \in \mathbf{F}^m$ given and $x \in \mathbf{F}^n$ unknown, the set of solutions is unchanged by elementary operations performed consistently upon A and b. The solutions may be read off from the RREF of $[A\ b]$. In fact, the RREF is unique, and two systems $Ax = b$ are *solution-equivalent* (have the same set of solutions) if and only if the two augmented matrices $[A\ b]$ have the same RREF.

We shall discuss the role of the RREF in rank and inversion later.

0.3.5 **Multiplicativity.** The most crucial and important property of the determinant function is that it is multiplicative: For $A, B \in M_n(\mathbf{F})$

$$\det AB = \det A \det B$$

This may be proved using elementary operations which row-reduce both A and B.

0.3.6 **Functional characterization of the determinant.** Thought of as a function of each row (or column) separately with the others fixed, the determinant is a linear function of the entries in the given row (or column). This is clear from the Laplace expansion: The coefficient of a given entry is just $\pm$ the complementary minor, which is fixed. A function that is linear with respect, in turn, to each of the subsets of a given partition of its arguments is called *multilinear*. This is a rather broad class. For example, the function $f(x_1, x_2) = x_1 x_2$ is multilinear, with the partition being $\{x_1\}$, $\{x_2\}$. Thus, the determinant is multilinear as a function of the entries of a matrix, with the partition corresponding to the rows (or columns).

It is natural to ask if any subset of the properties of the determinant noted thus far characterize it as a scalar-valued function of the n^2 entries of $A \in M_n$. The determinant is the unique function $f: M_n(\mathbf{F}) \to \mathbf{F}$ that is

(a) Multilinear;
(b) Alternating: the type 1 operation multiplies the result by -1; and
(c) Normalized: $f(I)=1$, where $I\in M_n(\mathbf{F})$ is the identity matrix.

The permanent function is also multilinear (as are other generalized matrix functions) and it is normalized, but it is not alternating.

0.4 Rank

An important nonnegative integer associated with each matrix $A\in M_{m,n}(\mathbf{F})$ is its *rank*, which we denote by rank A.

0.4.1 **Definition.** If $A\in M_{m,n}(\mathbf{F})$, rank A is the largest number of columns of A that constitute a linearly independent set. This set of columns is not, of course, unique, but the cardinality [number of elements] of this set is unique. It is a remarkable fact that rank $A^T=$ rank A. Therefore, rank may equivalently be defined in terms of linearly independent rows. Often this is phrased as "row rank = column rank."

0.4.2 **Rank and linear systems.** The linear system $Ax=b$ (0.3.4) may have 0, 1, or infinitely many solutions, but these are the only possibilities. If there is at least one solution, the system is said to be *consistent.* The linear system is consistent if and only if rank $[A\ b]=$ rank A. The m-by-$(n+1)$ matrix $[A\ b]$ is called the *augmented matrix,* and to say that the augmented matrix and the *coefficient matrix* A have the same rank is just to say that b is a linear combination of the columns of A. In this case, appending b to the columns of A does not increase the rank. A solution of the linear system $Ax=b$ is a vector x of coefficients which give b as a linear combination of the columns of A.

0.4.3 **RREF and rank.** Elementary operations do not change the rank of a matrix, and thus rank A is the same as the rank of the RREF of A, which is just the number of nonzero rows in the RREF. Calculation of the rank by calculation of the RREF suffers from *ill-conditioning:* Round-off errors in intermediate numerical calculations can make zero rows of the RREF appear to be nonzero, thereby affecting perception of the rank.

0.4.4 **Characterizations of rank.** The following statements about a given matrix $A\in M_{m,n}(\mathbf{F})$ are all equivalent; each can be useful in a different context.

(a) rank $A=k$;
(b) There exist k, and no more than k, rows of A that constitute a linearly independent set;
(c) There exist k, and no more than k, columns of A that constitute a linearly independent set;
(d) There is a k-by-k submatrix of A with nonzero determinant, but all $(k+1)$-by-$(k+1)$ submatrices of A have determinant 0;
(e) The dimension of the range of A is k;
(f) There is a set of k, but no more than k, linearly independent vectors b such that the linear system $Ax=b$ is consistent; and
(g) $k=n$ - the dimension of the null space of A.

0.4.5 **Rank inequalities.** Some fundamental inequalities involving the rank are the following.

(a) For $A\in M_{m,n}(\mathbf{F})$, rank $A\le\min\{m,n\}$.
(b) If rows and/or columns are deleted from a matrix, the rank of the resulting submatrix cannot be greater than the rank of the original matrix.
(c) If $A\in M_{m,k}(\mathbf{F})$ and $B\in M_{k,n}(\mathbf{F})$,

$$(\text{rank } A+\text{rank } B)-k\le \text{rank } AB\le\min\{\text{rank } A, \text{rank } B\}$$

(d) If $A,B\in M_{m,n}(\mathbf{F})$, $\text{rank}(A+B)\le\text{rank } A+\text{rank } B$.

Somewhat more subtle is an inequality of Frobenius, from which others follow:

(e) If $A\in M_{m,k}(\mathbf{F})$, $B\in M_{k,p}(\mathbf{F})$, and $C\in M_{p,n}(\mathbf{F})$, then

$$\text{rank } AB+\text{rank } BC\le\text{rank } B+\text{rank } ABC.$$

0.4.6 **Rank equalities.**

(a) If $A\in M_{m,n}(\mathbf{C})$, rank $A^*=$ rank $A^T=$ rank $\bar{A}=$ rank A.
(b) If $A\in M_m(\mathbf{F})$ and $C\in M_n(\mathbf{F})$ are nonsingular and $B\in M_{m,n}(\mathbf{F})$, then rank $AB=$ rank $B=$ rank $BC=$ rank ABC; that is, rank is unchanged upon left or right multiplication by a nonsingular matrix.
(c) If $A,B\in M_{m,n}(\mathbf{F})$, then rank $A=$ rank B if and only if there exist nonsingular $X\in M_m(\mathbf{F})$ and $Y\in M_n(\mathbf{F})$ such that $B=XAY$.
(d) If $A\in M_{m,n}(\mathbf{C})$, rank $A^*A=$ rank A.
(e) If $A\in M_{m,n}(\mathbf{F})$ has rank k, then

$$A=XBY$$

where $X \in M_{m,k}(\mathbf{F})$, $Y \in M_{k,n}(\mathbf{F})$, and $B \in M_k(\mathbf{F})$ is nonsingular. In particular, a matrix A that has rank 1 may always be written in the form $A = xy^T$ for some $x \in \mathbf{F}^m$, $y \in \mathbf{F}^n$.

0.5 Nonsingularity

A linear transformation or matrix is said to be *nonsingular* if it produces the output 0 only for the input 0. Otherwise, it is *singular*. If $A \in M_{m,n}(\mathbf{F})$ and $m < n$, then A is necessarily singular. If $A \in M_n(\mathbf{F})$, A is called *invertible* if there is a matrix $A^{-1} \in M_n(\mathbf{F})$ called the *inverse* of A such that $A^{-1}A = I$. Equivalently, A is invertible if the linear transformation A is one-to-one, and its inverse transformation (also linear) exists. If $A \in M_n$ and $A^{-1}A = I$, then $AA^{-1} = I$; A^{-1} is unique whenever it exists.

It is useful to be able to recognize in several different ways whether $A \in M_n(\mathbf{F})$ is nonsingular. The following are equivalent if $A \in M_n(\mathbf{F})$:

(a) A is nonsingular;
(b) A^{-1} exists;
(c) $\operatorname{rank} A = n$;
(d) The rows of A are linearly independent;
(e) The columns of A are linearly independent;
(f) $\det A \neq 0$;
(g) The dimension of the range of A is n;
(h) The dimension of the null space of A is 0;
(i) $Ax = b$ is consistent for each $b \in \mathbf{F}^n$;
(j) If $Ax = b$ is consistent, then the solution is unique;
(k) $Ax = b$ has a unique solution for each $b \in \mathbf{F}^n$;
(l) The only solution to $Ax = 0$ is $x = 0$; and
(m) 0 is not an eigenvalue of A (see Chapter 1).

The nonsingular matrices in $M_n(\mathbf{F})$ form a group, the *general linear group*, often denoted $GL(n, \mathbf{F})$.

0.6 The usual inner product

We adopt the convention of considering elements of $\mathbf{F}^n$ to be column vectors [i.e., $\mathbf{F}^n = M_{n,1}(\mathbf{F})$]. Thus if $x \in \mathbf{C}^n$, x^T and x^* are row vectors. Note that if $x \in \mathbf{R}^n$, $x^* = x^T$.

0.6.1 **Definition.** The scalar y^*x is an *inner product* (*scalar product*) of x and $y \in \mathbf{C}^n$ and is often denoted $\langle x, y\rangle \equiv y^*x$. Since it is possible to define inner products other than this one, we refer to this one as the *usual* or *standard inner product* on the vector space $\mathbf{C}^n$. Note that $\langle \cdot, \cdot \rangle$

is linear in the first argument ($\langle \alpha x_1 + \beta x_2, y\rangle = \alpha\langle x_1, y\rangle + \beta\langle x_2, y\rangle$ for all $\alpha, \beta \in \mathbf{C}$ and $x_1, x_2 \in \mathbf{C}^n$) and is *conjugate linear* in the second ($\langle x, \alpha y_1 + \beta y_2\rangle = \bar{\alpha}\langle x, y_1\rangle + \bar{\beta}\langle x, y_2\rangle$ for all $\alpha, \beta \in \mathbf{C}$ and $y_1, y_2 \in \mathbf{C}^n$).

0.6.2 **Orthogonality.** Two vectors $x, y \in \mathbf{C}^n$ are called *orthogonal* if $\langle y, x\rangle = 0$. In two and three dimensions, this has the conventional geometric interpretation of perpendicular. A set of vectors $\{x_1, \dots, x_k\} \subset \mathbf{C}^n$ is said to be orthogonal if each pair of vectors in the set is orthogonal. A set of orthogonal vectors, none of which is the zero vector, is necessarily linearly independent.

0.6.3 **The Cauchy-Schwarz inequality.** If $x \in \mathbf{C}^n$, the nonnegative scalar $\langle x, x\rangle^{1/2}$ is the *Euclidean length* of x. A vector whose Euclidean length is 1 is said to be *normalized* (or, sometimes, a *unit vector*). For any nonzero vector $x \in \mathbf{C}^n$, $x/\langle x, x\rangle^{1/2}$ is a normalized vector which points in the same direction as x. The fundamental Cauchy-Schwarz inequality states that

$$|\langle y, x\rangle| \leq \langle x, x\rangle^{1/2}\langle y, y\rangle^{1/2}$$

for all $x, y \in \mathbf{C}^n$ with equality if and only if x and y are dependent. Generalizing the notion of orthogonality, the *angle* θ between two nonzero vectors $x, y \in \mathbf{C}^n$ may be defined unambiguously by

$$\cos\theta = \frac{|\langle y, x\rangle|}{\langle x, x\rangle^{1/2}\langle y, y\rangle^{1/2}}, \qquad 0 \leq \theta \leq \frac{\pi}{2}$$

0.6.4 **Gram-Schmidt orthonormalization.** It is intuitively plausible that a set of linearly independent vectors (which form a basis for their span) may be replaced by an *orthonormal* (orthogonal and individually normalized) basis for the same space. Although this replacement may, in principle, be carried out in infinitely many ways, there is a very simple and far-reaching algorithm for carrying out this replacement: the *Gram-Schmidt (orthonormalization) process*. Let $\{x_1, \dots, x_n\}$ be a set of n linearly independent vectors in a complex vector space and let $\{z_1, \dots, z_n\}$ be the orthogonal set of normalized vectors to be determined. The z_i may be calculated in turn as follows. Let $y_1 = x_1$ and choose

$$z_1 = \frac{y_1}{\langle y_1, y_1\rangle^{1/2}}$$

so that z_1 is normalized. Let $y_2 = x_2 - \langle x_2, z_1\rangle z_1$, so that y_2 is orthogonal to z_1, and choose

$$z_2 = \frac{y_2}{\langle y_2, y_2 \rangle^{1/2}}$$

so that z_2 is normalized and is orthogonal to z_1. The process continues similarly. Assuming $z_1, \ldots, z_{k-1}$ have been determined, let

$$y_k = x_k - \langle x_k, z_{k-1} \rangle z_{k-1} - \langle x_k, z_{k-2} \rangle z_{k-2} - \cdots - \langle x_k, z_1 \rangle z_1$$

so that y_k is orthogonal to $z_1, \ldots, z_{k-1}$, and again normalize y_k to get

$$z_k = \frac{y_k}{\langle y_k, y_k \rangle^{1/2}}$$

Continue until the desired orthonormal vectors $z_1, \ldots, z_n$ have been produced. Note that an infinite orthonormal set could be produced from a countably infinite linearly independent set in an infinite-dimensional vector space in this way.

At each step in the Gram–Schmidt process, the orthonormal vectors $z_1, \ldots, z_k$ are a linear combination of the original independent vectors $x_1, \ldots, x_k$ only (and vice versa). If we denote $Z = [z_1 \, z_2 \cdots z_n]$ and $X = [x_1 \, x_2 \cdots x_n]$, matrices that have as columns the vectors z_i and x_i, respectively, then $Z = XR$, where the matrix $R = [r_{ij}]$ is nonsingular and upper triangular; that is, $r_{ij} = 0$ whenever $i > j$.

Finally, we note that the Gram-Schmidt process may be applied to any finite or countable (not necessarily linearly independent) sequence of vectors. If the set is not independent, it will produce a vector $y_k = 0$ for the least value of k for which $\{x_1, \ldots, x_k\}$ is a linearly dependent set. In this case, x_k is a linear combination of $x_1, \ldots, x_{k-1}$. Substitution of x_{k+1} for x_k and continuation of the Gram-Schmidt process can answer such questions as: What is a basis for, or the dimension of, the span of $\{x_1, \ldots, x_n\}$?

0.6.5 **Orthonormal bases.** An orthonormal set of vectors is just an orthogonal set of vectors, each of which is normalized. Such a set cannot contain the vector 0 and is necessarily linearly independent. An orthonormal basis is a basis whose vectors constitute an orthonormal set. Since any basis may be transformed to an orthonormal basis (via Gram–Schmidt), any finite-dimensional complex vector space has an orthonormal basis. Such a basis is pleasant to work with, since the cross-terms in inner product calculations all vanish.

0.6.6 **Orthogonal complements.** Given any subset $S \subset \mathbf{C}^n$, the *orthogonal complement* of S is the set

$$S^\perp \equiv \{x \in \mathbf{C}^n ; x^* y = 0 \quad \text{for all} \quad y \in S\}$$

Even if S is not a subspace, $S^\perp$ is always a subspace. We have $(S^\perp)^\perp = \operatorname{Span} S$, and $(S^\perp)^\perp = S$ if S is a subspace. It is always the case that $\dim S^\perp + \dim (S^\perp)^\perp = n$. In the context of the linear system $Ax = b$, $A \in M_{m,n}$, it is worth noting that the range of A is just the orthogonal complement of the null space of A^*; that is, $Ax = b$ has a solution (not necessarily unique) if and only if $b^* z = 0$ for all $z \in \mathbf{C}^m$ such that $A^* z = 0$.

0.7 Partitioned matrices

Analogous to a partition of a set, a partition of a matrix is an exhaustive decomposition of the matrix into mutually exclusive submatrices such that each entry of the original matrix falls in one and only one submatrix of the partition. Partitioning of matrices is often a convenient device for perception of useful structure.

0.7.1 **Submatrices.** Let $A \in M_{m,n}(\mathbf{F})$. For index sets $\alpha \subseteq \{1, \ldots, m\}$ and $\beta \subseteq \{1, \ldots, n\}$, we denote the (sub)matrix that lies in the rows of A indexed by α and the columns indexed by β as $A(\alpha, \beta)$. For example,

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} (\{1,3\}, \{1,2,3\}) = \begin{bmatrix} 1 & 2 & 3 \\ 7 & 8 & 9 \end{bmatrix}$$

If $m = n$ and $\beta = \alpha$, the submatrix $A(\alpha, \alpha)$ is called a *principal submatrix* of A and is abbreviated $A(\alpha)$. Often it is convenient to indicate a submatrix or principal submatrix via deletion, rather than inclusion, of rows or columns. This may be accomplished by complementing the index sets. For example, $A(\alpha', \beta')$ is the result of *deleting* the rows indicated by α and the columns indicated by β.

The determinant of a square submatrix of the matrix A is called a *minor* of A. If the submatrix is a principal submatrix, then the minor is a *principal minor.* A signed minor, such as those appearing in the Laplace expansion (0.3.1) $[(-1)^{i+j} \det A_{ij}]$ is called a *cofactor* of A. By convention, the empty principal minor is 1; that is, $\det A(\phi) = 1$.

0.7.2 **Partitions and multiplication.** If $\alpha_1, \ldots, \alpha_t$ constitute a partition of $\{1, \ldots, m\}$ and $\beta_1, \ldots, \beta_s$ constitute a partition of $\{1, \ldots, n\}$, then the matrices $A(\alpha_i, \beta_j)$ form a *partition* of the matrix $A \in M_{m,n}(\mathbf{F})$, $1 \le i \le t$, $1 \le j \le s$. If $A \in M_{m,n}(\mathbf{F})$ and $B \in M_{n,p}(\mathbf{F})$ are partitioned so that the two partitions of $\{1, \ldots, n\}$ coincide, the two matrix partitions are said to be *conformal.* In this event,

$$[AB](\alpha_i, \gamma_j) = \sum_{k=1}^{s} A(\alpha_i, \beta_k) B(\beta_k, \gamma_j)$$

where $A(\alpha_i, \beta_k)$ and $B(\beta_k, \gamma_j)$ are conformal partitions of A and B. The left-hand side is a submatrix of the product AB (calculated in the usual way), and each summand on the right is a standard matrix product. Thus, multiplication of conformally partitioned matrices mimics usual matrix multiplication. Addition of partitioned matrices also makes sense when the summands are partitioned identically.

0.7.3 **The inverse of a partitioned matrix.** It is sometimes useful to know the corresponding blocks in the inverse of a partitioned, nonsingular matrix A, that is, to present the inverse of a partitioned matrix in correspondingly partitioned form. This may be done in a variety of apparently different, but equivalent, ways - assuming that certain submatrices of $A \in M_n(\mathbf{F})$ and A^{-1} are also nonsingular. For simplicity, let A be partitioned as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

with $A_{ii} \in M_{n_i}(\mathbf{F})$, $i = 1, 2$ and $n_1 + n_2 = n$. A useful expression for the correspondingly partitioned presentation of A^{-1} is

$$\begin{bmatrix} [A_{11} - A_{12}A_{22}^{-1}A_{21}]^{-1} & A_{11}^{-1}A_{12}[A_{21}A_{11}^{-1}A_{12} - A_{22}]^{-1} \\ [A_{21}A_{11}^{-1}A_{12} - A_{22}]^{-1}A_{21}A_{11}^{-1} & [A_{22} - A_{21}A_{11}^{-1}A_{12}]^{-1} \end{bmatrix}$$

assuming that all the relevant inverses exist. Or, in general index set notation, we may write

$$A^{-1}(\alpha) = [A(\alpha) - A(\alpha, \alpha')A(\alpha')^{-1}A(\alpha', \alpha)]^{-1}$$

and

$$A^{-1}(\alpha, \alpha') = A(\alpha)^{-1}A(\alpha, \alpha')[A(\alpha', \alpha)A(\alpha)^{-1}A(\alpha, \alpha') - A(\alpha')]^{-1}$$

again assuming that the relevant inverses exist. Other presentations are possible. Note that $A^{-1}(\alpha)$ is a submatrix of A^{-1}, while $A(\alpha)^{-1}$ is the inverse of a submatrix of A, and these two objects are not, in general, the same.

0.7.4 **The inverse of a small-rank adjustment.** If the inverse of a given matrix is known, it is also often useful to know how the inverse changes upon addition of a matrix of "small" rank. There is a simple formula that, if the form of the adjustment matrix is sufficiently simple, can make the new inverse simpler to compute than starting from scratch. Suppose a nonsingular matrix $A \in M_n(\mathbf{F})$ has a known inverse A^{-1} and consider

$$B = A + XRY$$

where X is n-by-r, Y is r-by-n, and R is r-by-r and nonsingular. If B is nonsingular, then

$$B^{-1} = A^{-1} - A^{-1}X(R^{-1} + YA^{-1}X)^{-1}YA^{-1}$$

If r is much smaller than n, then R and $R^{-1} + YA^{-1}X$ may be much easier to invert than B, and if A is easy to invert and has a form that renders the multiplications simple, then use of this formula may be competitive with direct inversion of B. For example, if the adjustment has rank 1, X is n-by-1, Y is 1-by-n, and $R = [1]$, the formula becomes

$$B^{-1} = A^{-1} - \frac{1}{1 + YA^{-1}X}A^{-1}XYA^{-1}$$

(note that $XY = B - A$ in this case). In particular, if

$$B = I + xy^T$$

for $x, y \in \mathbf{F}^n$, $I \in M_n(\mathbf{F})$, then

$$B^{-1} = I - \frac{1}{1 + y^Tx}xy^T$$

if $y^Tx \neq -1$.

0.8 Determinants again

Some additional facts about and identities for the determinant are useful for reference. Most of these are not generally found in elementary treatments.

0.8.1 **Compound matrices.** The array of all minors of a given size from a given matrix $A \in M_{m,n}(\mathbf{F})$ is called a *compound matrix* of A. In particular, the $\binom{m}{k}$-by-$\binom{n}{k}$ matrix whose α, β entry is $\det A(\alpha, \beta)$ is called the *kth compound matrix* of A and is denoted by $C_k(A)$. Here, $\alpha \subseteq \{1, \ldots, m\}$ and $\beta \subseteq \{1, \ldots, n\}$ are index sets of cardinality $k \leq \min\{m, n\}$, usually ordered lexicographically, that is, $\{1, 2, 4\}$ before $\{1, 2, 5\}$ before $\{1, 3, 4\}$ and so on. For example, if

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

then

$$C_2(A) = \begin{bmatrix} \det\begin{bmatrix}1&2\\4&5\end{bmatrix} & \det\begin{bmatrix}1&3\\4&6\end{bmatrix} & \det\begin{bmatrix}2&3\\5&6\end{bmatrix} \\ \det\begin{bmatrix}1&2\\7&8\end{bmatrix} & \det\begin{bmatrix}1&3\\7&9\end{bmatrix} & \det\begin{bmatrix}2&3\\8&9\end{bmatrix} \\ \det\begin{bmatrix}4&5\\7&8\end{bmatrix} & \det\begin{bmatrix}4&6\\7&9\end{bmatrix} & \det\begin{bmatrix}5&6\\8&9\end{bmatrix} \end{bmatrix} = \begin{bmatrix} -3 & -6 & -3 \\ -6 & -12 & -6 \\ -3 & -6 & -3 \end{bmatrix}$$

If $A \in M_{m,k}(\mathbf{F})$ and $B \in M_{k,n}(\mathbf{F})$, then

$$C_r(AB) = C_r(A)C_r(B), \qquad r \le \min\{m, k, n\}$$

Also,

$$C_r(tA) = t^r C_r(A), \qquad t \in \mathbf{F}$$

$$\text{If } I \in M_n, \qquad C_k(I) = I \in M_{\binom{n}{k}}$$

$$\text{If } A \in M_n \text{ is nonsingular}, \qquad C_k(A)^{-1} = C_k(A^{-1})$$

$$\text{If } A \in M_{m,n}(\mathbf{F}), \qquad C_k(A^T) = C_k(A)^T$$

and

$$\text{If } A \in M_{m,n}(\mathbf{C}), \qquad C_k(A^*) = C_k(A)^*$$

0.8.2 **The classical adjoint and the inverse.** If $A \in M_n(\mathbf{F})$, the transposed matrix of cofactors $B = [b_{ij}] \in M_n(\mathbf{F})$ defined by

$$b_{ij} = (-1)^{i+j} \det A(\{j\}', \{i\}')$$

is called the *(classical) adjoint* of A and is often denoted adj A. The term *adjugate* is sometimes used in place of *adjoint* to avoid confusion with the Hermitian adjoint A^*. Note that

$$\operatorname{adj} A = EC_{n-1}(A)^T E$$

where

$$E = \begin{bmatrix} 1 & & & & & 0 \\ & -1 & & & & \\ & & 1 & & & \\ & & & -1 & & \\ & 0 & & & \ddots & \\ & & & & & \pm 1 \end{bmatrix}$$

A calculation using the Laplace expansion for the determinant shows that

$$(\operatorname{adj} A)A = A(\operatorname{adj} A) = (\det A)I$$

Thus, if A is nonsingular ($\det A \ne 0$),

$$A^{-1} = \frac{1}{\det A} \operatorname{adj} A$$

Use of the adjoint is generally not a good way to calculate the inverse of a matrix numerically, but the adjoint can be useful as an analytic way to present the inverse.

0.8.3 **Cramer's rule.** Cramer's rule is one method of presenting the unique solution to the linear system $Ax = b$ when $A \in M_n(\mathbf{F})$ is nonsingular. It bears the same computational caveats as the adjoint presentation of the inverse, and it is generally useful only when there is a need to present analytically a particular component of the solution. If x_i is the ith component of the solution vector $x \in \mathbf{F}^n$, then Cramer's rule states that

$$x_i = \frac{\det(A \underset{i}{\leftarrow} b)}{\det A}$$

The notation $A \underset{i}{\leftarrow} b$ denotes that matrix in M_n whose ith column is b and whose remaining columns coincide with those of A. Cramer's rule follows directly from the multiplicativity of the determinant. The system $Ax = b$ may be rewritten as

$$A(I \underset{i}{\leftarrow} x) = A \underset{i}{\leftarrow} b$$

and taking determinants of both sides (using multiplicativity) gives

$$(\det A)\det(I \underset{i}{\leftarrow} x) = \det(A \underset{i}{\leftarrow} b)$$

But $\det(I \underset{i}{\leftarrow} x) = x_i$, and the formula follows.

0.8.4 **Minors of the inverse.** An important fact, which generalizes the adjoint formula for the inverse of a nonsingular matrix, and which relates the minors of A^{-1} to those of $A \in M_n(\mathbf{F})$, is the following:

$$\det A^{-1}(\alpha', \beta') = (-1)^{(\sum_{i \in \alpha} i + \sum_{j \in \beta} j)} \frac{\det A(\beta, \alpha)}{\det A}$$

For principal submatrices, this formula assumes the simple form

$$\det A^{-1}(\alpha') = \frac{\det A(\alpha)}{\det A}$$

0.8.5 **Schur complements and determinantal formulae.** Let $\alpha \subseteq \{1, \ldots, n\}$ be an index set such that $A(\alpha)$ is nonsingular for a given matrix $A \in M_n(\mathbf{F})$. Denote the inverse of $A(\alpha)$ by $A(\alpha)^{-1}$. An important formula for $\det A$, based upon the 2-partition of A using α and α', is

$$\det A = \det A(\alpha) \det[A(\alpha') - A(\alpha', \alpha)A(\alpha)^{-1}A(\alpha, \alpha')]$$

Notice that this generalizes the familiar formula for the determinant of a 2-by-2 matrix displayed in (0.3.1). The special matrix

$$A(\alpha') - A(\alpha', \alpha)A(\alpha)^{-1}A(\alpha, \alpha')$$

is called the *Schur complement* of $A(\alpha)$ in A. The Schur complement formula for $\det A$ may be verified via multiplication of

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \text{ by } \begin{bmatrix} I & -A_{11}^{-1} & A_{12} \\ 0 & I & \end{bmatrix}$$

and identification of A_{11} with $A(\alpha)$. Note that the Schur complement has already risen in the partitioned form for A^{-1} (0.7.3).

0.8.6 **Sylvester's identity.** Let $\alpha \subseteq \{1, \ldots, n\}$ be a fixed index set and let $B = [b_{ij}] \in M_{n-k}(\mathbf{F})$ be defined by

$$b_{ij} = \det A(\alpha \cup \{i\},\ \alpha \cup \{j\})$$

where k is the cardinality of α, $i, j \in \{1, \ldots, n\}$ are indices *not* contained in α, and $A \in M_n(\mathbf{F})$. Another useful determinantal identity is

$$\det B = [\det A(\alpha)]^{n-k-1} \det A$$

0.8.7 **Cauchy–Binet formula.** This useful formula can be remembered because of its similarity in appearance to the formula for matrix multiplication. This is no accident, since it is equivalent to the multiplicativity of the compound matrix (0.8.1). Let $A \in M_{m,k}(\mathbf{F})$, $B \in M_{k,n}(\mathbf{F})$, and $C = AB$. Further, let $1 \le r \le \min\{m, k, n\}$, and let $\alpha \subseteq \{1, \ldots, m\}$ and $\beta \subseteq \{1, \ldots, n\}$ be index sets, each of cardinality r. An expression for the α, β minor of C is

$$\det C(\alpha, \beta) = \sum_{\gamma} \det A(\alpha, \gamma) \det B(\gamma, \beta)$$

where the sum is taken over all index sets $\gamma \subseteq \{1, \ldots, k\}$ of cardinality r.

0.8.8 **Relations among minors.** Let $A \in M_{m,n}(\mathbf{F})$ be given and let a fixed index set $\alpha \subseteq \{1, \ldots, m\}$ of cardinality k be given. The minors

$$\det A(\alpha, \omega),$$

as $\omega \subseteq \{1, \ldots, n\}$ runs over *ordered* index sets of cardinality k, are not algebraically independent since there are more minors than there are distinct entries among the submatrices. Quadratic relations are known

among these minors. Let $i_1, i_2, \ldots, i_k \in \{1, \ldots, n\}$ be k distinct indices, not necessarily in natural order, and let

$$A(\alpha; i_1, \ldots, i_k)$$

denote the matrix whose rows are indicated by α and whose jth column is column i_j of $A(\alpha, \{1, \ldots, n\})$. The difference between this and our prior notation is that columns may not occur in natural order as in $A(\{1, 3\}; 4, 2)$, whose first column has the 1,4 and 3,4 entries of A. We then have the relations

$$\det A(\alpha; i_1, \ldots, i_k) \det A(\alpha; j_1, \ldots, j_k)$$
$$= \sum_{t=1}^{k} \det A(\alpha; i_1, \ldots, i_{s-1}, j_t, i_{s+1}, \ldots, i_k) \det A(\alpha; j_1, \ldots, j_{t-1}, i_s, j_{t+1}, \ldots, j_k)$$

for each $s = 1, \ldots, k$ and all sequences of distinct indices

$$i_1, \ldots, i_k \in \{1, \ldots, n\} \quad \text{and} \quad j_1, \ldots, j_k \in \{1, \ldots, n\}$$

0.9 Special types of matrices

Certain matrices of special form arise frequently and have important properties. Some of these are worth cataloging here for reference and terminology.

0.9.1 **Diagonal matrices.** The matrix $D = [d_{ij}] \in M_n$ is called *diagonal* if $d_{ij} = 0$ whenever $j \ne i$. Conventionally, we denote such a matrix as $D = \text{diag}(d_{11}, \ldots, d_{nn})$ or $D = \text{diag}\, d$, where d is the vector of diagonal entries of D. If all the diagonal entries of a diagonal matrix are positive (nonnegative) real numbers, we refer to it as a *positive* (*nonnegative*) *diagonal matrix*. Note that the term *positive diagonal matrix* means that the matrix is diagonal in addition to having positive diagonal entries; it does not refer to a general matrix all of whose diagonal entries happen to be positive. The identity matrix $I \in M_n$ is an example of a positive diagonal matrix. A diagonal matrix $D \in M_n$ is called a *scalar matrix* if the diagonal entries of D are all equal; that is, $D = \alpha I$ for some $\alpha \in \mathbf{C}$. Left or right multiplication of a matrix by a scalar matrix has the same effect as multiplying it by the corresponding scalar.

The determinant of a diagonal matrix is just the product of its diagonal entries: $\det D = \prod_{i=1}^{n} d_{ii}$. Thus, a diagonal matrix is nonsingular if and only if all its diagonal entries are nonzero. *Left* multiplication of $A \in M_n$ by a diagonal matrix D, that is, DA, multiplies the *rows* of A by the diagonal entries of D (the ith row of A is multiplied by d_{ii}, $i = 1, \ldots, n$). *Right* multiplication by D, that is, AD, multiplies the *columns* of A by

the diagonal entries of D. Thus, all diagonal matrices commute with each other under multiplication, and a diagonal matrix D commutes with a given matrix $A=[a_{ij}]\in M_n$ if and only if $a_{ij}=0$ whenever the ith and jth diagonal entries of D differ. The product of two diagonal matrices is just the diagonal matrix of pair-wise products of their respective diagonal entries and similarly for positive integer powers of a single diagonal matrix.

0.9.2 **Block diagonal matrices.** A matrix $A\in M_n$ of the form

$$A=\begin{bmatrix} A_{11} & & & 0 \\ & A_{22} & & \\ & & \ddots & \\ 0 & & & A_{kk} \end{bmatrix}$$

in which $A_{ii}\in M_{n_i}$, $i=1,\ldots,k$, and $\sum_{i=1}^k n_i=n$, is called *block diagonal.* Notationally, such a matrix is often indicated as $A=A_{11}\oplus A_{22}\oplus\cdots\oplus A_{kk}$ or, more briefly, $\oplus\sum_{i=1}^k A_{ii}$; this is called the *direct sum* of the matrices $A_{11},\ldots,A_{kk}$. Thinking in terms of partitioned multiplication, many properties of block diagonal matrices generalize those of diagonal matrices. For example, $\det(\oplus\sum_{i=1}^k A_{ii})=\prod_{i=1}^k \det A_{ii}$, so that $A=\oplus\sum A_{ii}$ is nonsingular if and only if each A_{ii} is nonsingular, $i=1,\ldots,k$. Furthermore, two direct sums $A=\oplus\sum_{i=1}^k A_{ii}$ and $B=\oplus\sum_{i=1}^k B_{ii}$, in which each pair A_{ii}, B_{ii} is the same size, commute if and only if A_{ii} and B_{ii} commute, $i=1,\ldots,k$. Also, $\operatorname{rank}(\oplus\sum_{i=1}^k A_{ii})=\sum_{i=1}^k \operatorname{rank} A_{ii}$.

0.9.3 **Triangular matrices.** The matrix $T=[t_{ij}]\in M_n$ is said to be *upper triangular* if $t_{ij}=0$ whenever $j<i$. If $t_{ij}=0$ whenever $j\le i$, then T is said to be *strictly upper triangular.* Analogously, T is said to be *lower triangular* (or *strictly lower triangular*) if its transpose is upper triangular (or strictly upper triangular). Triangular matrices are like diagonal matrices in that the determinant of a triangular matrix is the product of its diagonal entries. However, triangular matrices (of either sort) do not necessarily commute with other triangular matrices. Left multiplication of $A\in M_n$ by a lower triangular matrix L, that is, LA, replaces the ith row of A by a linear combination of the first through ith rows of A. Sometimes the terms *right* (in place of *upper*) and *left* (in place of *lower*) are used in reference to triangular matrices. The rank of a triangular matrix is at least, and can be greater than, the number of nonzero entries on the main diagonal.

0.9.4 **Block triangular matrices.** A matrix $A\in M_n$ of the form

$$A=\begin{bmatrix} A_{11} & & & * \\ & A_{22} & & \\ & & \ddots & \\ 0 & & & A_{kk} \end{bmatrix}$$

in which $A_{ii}\in M_{n_i}$, $i=1,\ldots,k$, $\sum_{i=1}^k n_i=n$, and "$*$" denotes any entry, is called *block upper triangular. Block lower triangular, strictly block lower triangular,* and *strictly block upper triangular* may be defined similarly. The determinant of a block triangular matrix is the product of the determinants of the diagonal blocks. The rank of a block triangular matrix is at least, and can be greater than, the sum of the ranks of the diagonal blocks.

0.9.5 **Permutation matrices.** A matrix $P\in M_n$ is called a *permutation matrix* if exactly one entry in each row and column is equal to 1, and all other entries are 0. Multiplication by such matrices effects a permutation of the rows or columns of the object multiplied. For example,

$$P=\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}\in M_3$$

is a permutation matrix, and

$$P\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}=\begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$$

is a permutation of the rows (components) of the vector $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$, namely, the permutation that sends the first item to the second position, sends the second item to the first position, and leaves the third item in the third position. In general, left multiplication of a matrix $A\in M_{m,n}$ by a permutation matrix $P\in M_m$ permutes the rows of A, while right multiplication of a matrix $A\in M_{m,n}$ by a permutation matrix $P\in M_n$ permutes the columns of A. The matrix that carries out a type 1 elementary operation (0.3.3) is an example of a special type of permutation matrix called a *transposition.*

The determinant of a permutation matrix is ± 1 [exactly one summand in the formula of (0.3.2) is nonzero], so that permutation matrices are

necessarily nonsingular. Although permutation matrices do not, in general, commute under multiplication, the product of two permutation matrices is again a permutation matrix. Since the identity is a permutation matrix and $P^T = P^{-1}$ for every permutation P, the permutation matrices constitute a subgroup of $GL(n, \mathbf{C})$, the group of nonsingular matrices in M_n, which has finite cardinality $n!$. In fact, any permutation matrix is a product of transpositions.

Since $P^T = P^{-1}$ permutes columns in the same way that the permutation matrix $P \in M_n$ permutes rows, the transformation $A \to PAP^T$ permutes the rows and columns of $A \in M_n$ in the same way. In the context of linear equations with coefficient matrix A, this transformation amounts to renumbering the variables. A matrix $A \in M_n$ such that PAP^T is triangular for some permutation matrix P is called *essentially triangular*. These matrices have much in common with triangular matrices.

0.9.6 **Circulant matrices.** A matrix $A \in M_n$ of the form

$$A = \begin{bmatrix} a_1 & a_2 & \cdots\cdots & & a_n \\ a_n & a_1 & a_2 & \cdots & a_{n-1} \\ a_{n-1} & a_n & a_1 & \cdots & a_{n-2} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ a_2 & a_3 & \cdots & a_n & a_1 \end{bmatrix}$$

is called a *circulant matrix*. Each row is just the previous row cycled forward one step, so that the entries in each row are just a cyclic permutation of those in the first. The permutation matrix

$$C = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & 1 & & \vdots \\ & & & \ddots & 0 \\ 0 & & & \ddots & 1 \\ 1 & 0 & & \cdots & 0 \end{bmatrix}$$

is called the *basic circulant permutation* matrix. A matrix $A \in M_n$ can be written in the form

$$A = \sum_{k=0}^{n-1} a_{k+1} C^k$$

if and only if A is a circulant. Here $C^0 \equiv I \equiv C^n$, and the coefficients $a_1, a_2, \ldots, a_n$ are just the entries of the first row of A. Because of this representation, circulant matrices have very nice structure, which can be related to C. Since $C^n = I$, the product of two circulants is again a circulant. Furthermore, circulants commute under multiplication. There are generalizations in which, for example, the rows are cycled forward (or backward) a fixed number of steps that is greater than 1.

0.9.7 **Toeplitz matrices.** A matrix $A = [a_{ij}] \in M_{n+1}$ of the form

$$A = \begin{bmatrix} a_0 & a_1 & a_2 & \cdots\cdots & a_n \\ a_{-1} & a_0 & a_1 & \cdots\cdots & a_{n-1} \\ a_{-2} & a_{-1} & a_0 \quad a_1 & \cdots & \vdots \\ \vdots & & \ddots & \ddots & a_1 \\ a_{-n} & a_{-n+1} & \cdots & a_{-1} & a_0 \end{bmatrix}$$

is called a *Toeplitz matrix*. The general term $a_{ij} = a_{j-i}$ for some given sequence $a_{-n}, a_{-n+1}, \ldots, a_{-1}, a_0, a_1, a_2, \ldots, a_{n-1}, a_n \in \mathbf{C}$. The entries of A are constant down the diagonals parallel to the main diagonal. The Toeplitz matrices

$$B = \begin{bmatrix} 0 & 1 & & \mathbf{0} \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ \mathbf{0} & & & 0 \end{bmatrix} \quad \text{and} \quad F = \begin{bmatrix} 0 & & & \mathbf{0} \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ \mathbf{0} & & 1 & 0 \end{bmatrix}$$

are called the "backward shift" and "forward shift" because of their effect on the elements of the standard basis $\{e_1, \ldots, e_{n+1}\}$. A matrix $A \in M_{n+1}$ can be written in the form

$$A = \sum_{k=1}^{n} a_{-k} F^k + \sum_{k=0}^{n} a_k B^k$$

if and only if A is a Toeplitz matrix. Toeplitz matrices arise naturally in problems involving trigonometric moments.

0.9.8 **Hankel matrices.** A matrix $A \in M_{n+1}$ of the form

$$A = \begin{bmatrix} a_0 & a_1 & a_2 & \cdots\cdots & & a_n \\ a_1 & a_2 & a_3 & \cdots\cdots & a_n & a_{n+1} \\ a_2 & a_3 & & \ddots & & a_{n+2} \\ \vdots & \cdot & & & \ddots & \vdots \\ & a_n & & & & \\ a_n & a_{n+1} & a_{n+2} & \cdots\cdots & & a_{2n} \end{bmatrix}$$

is called a *Hankel matrix*. The general term $a_{ij} = a_{i+j-2}$ for some given sequence $a_0, a_1, a_2, \ldots, a_{2n-1}, a_{2n}$. The entries of A are constant along the diagonals perpendicular to the main diagonal. Hankel matrices arise naturally in problems involving power moments. Notice that if

$$P = \begin{bmatrix} 0 & & & 1 \\ & & 1 & \\ & \cdot^{\cdot^{\cdot}} & & \\ & 1 & & \\ 1 & & & 0 \end{bmatrix},$$

the "backward identity" permutation, then PT is a Hankel matrix for any Toeplitz matrix T, and PH is a Toeplitz matrix for any Hankel matrix H. Since $P = P^T = P^{-1}$ and Hankel matrices are symmetric, this means that any Toeplitz matrix is a product of two symmetric matrices (P and a Hankel matrix).

0.9.9 **Hessenberg matrices.** The matrix $A = [a_{ij}] \in M_n$ is said to be in *upper Hessenberg form* or to be an *upper Hessenberg matrix* if $a_{ij} = 0$ for $i > j+1$:

$$A = \begin{bmatrix} a_{11} & a_{12} & & \cdots & & a_{1n} \\ a_{21} & a_{22} & & & & \\ 0 & a_{32} & \ddots & & & \\ \vdots & 0 & \ddots & & & \\ & \vdots & \ddots & & & \\ 0 & 0 & \cdots & 0 & a_{n,n-1} & a_{nn} \end{bmatrix}$$

The matrix $A \in M_n$ is called *lower Hessenberg* if A^T is upper Hessenberg.

0.9.10 **Tridiagonal matrices.** A matrix $A = [a_{ij}] \in M_n$ that is *both* upper and lower Hessenberg is called *tridiagonal*, that is, A is tridiagonal if $a_{ij} = 0$, whenever $|i-j| > 1$:

$$A = \begin{bmatrix} a_{11} & a_{12} & & & 0 \\ a_{21} & a_{22} & \ddots & & \\ & a_{32} & \ddots & \ddots & \\ & & \ddots & \ddots & a_{n-1,n} \\ 0 & & & a_{n,n-1} & a_{n,n} \end{bmatrix}$$

The determinant of a tridiagonal matrix is easy to calculate inductively. Note that

$$\det A(\{1, 2, \ldots, k+1\})$$
$$= a_{k+1,k+1} \det A(\{1, \ldots, k\}) - a_{k+1,k} a_{k,k+1} \det A(\{1, \ldots, k-1\}),$$
$$k = 2, \ldots, n-1$$

0.9.11 **Matrices and Lagrange interpolation.** A *Vandermonde matrix* $A \in M_n(\mathbf{F})$ is a matrix of the form

$$A = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & x_2^3 & \cdots & x_2^{n-1} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & x_n^3 & \cdots & x_n^{n-1} \end{bmatrix} \tag{0.9.11.1}$$

where $x_1, x_2, \ldots, x_n \in \mathbf{F}$; that is, $A = [a_{ij}]$ with $a_{ij} = x_i^{j-1}$. It is a fact that

$$\det A = \prod_{\substack{i,j=1 \\ i>j}}^{n} (x_i - x_j) \tag{0.9.11.2}$$

so a Vandermonde matrix is nonsingular if and only if the n parameters $x_1, x_2, \ldots, x_n$ are distinct.

The Vandermonde matrix arises in the *interpolation problem* of finding a polynomial $p(x) = a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1 x + a_0$ of degree at most $n-1$ with coefficients from the field $\mathbf{F}$ such that

$$\begin{aligned} p(x_1) &= a_0 + a_1 x_1 + a_2 x_1^2 + \cdots + a_{n-1} x_1^{n-1} = y_1 \\ p(x_2) &= a_0 + a_1 x_2 + a_2 x_2^2 + \cdots + a_{n-1} x_2^{n-1} = y_2 \\ &\vdots \\ p(x_n) &= a_0 + a_1 x_n + a_2 x_n^2 + \cdots + a_{n-1} x_n^{n-1} = y_n \end{aligned} \tag{0.9.11.3}$$

where $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ are given elements of $\mathbf{F}$. The interpolation conditions (0.9.11.3) are a system of n equations for the n unknown coefficients $a_0, a_1, \ldots, a_{n-1}$, and they have the form $Aa = y$, where $a = [a_0, a_1, \ldots, a_{n-1}]^T \in \mathbf{F}^n$, $y = [y_1, y_2, \ldots, y_n]^T \in \mathbf{F}^n$, and $A \in M_n(\mathbf{F})$ is the Vandermonde matrix (0.9.11.1). This interpolation problem always has a solution if the points $x_1, x_2, \ldots, x_n$ are distinct, for A is nonsingular in this event.

If the points $x_1, x_2, \ldots, x_n$ are distinct, the coefficients of the interpolating polynomial could in principle be obtained by solving the system (0.9.11.3), but it is usually more useful to represent the interpolating polynomial $p(x)$ in terms of the *special Lagrange interpolating polynomials*

$$L_i(x) = \frac{\prod_{\substack{j=1 \\ j \neq i}}^{n} (x - x_j)}{\prod_{\substack{j=1 \\ j \neq i}}^{n} (x_i - x_j)}, \qquad i = 1, 2, \ldots, n$$

Each polynomial $L_i(x)$ has degree $n-1$ and has the property that $L_i(x_k) = 0$ if $k \neq i$, but $L_i(x_i) = 1$. Thus, we have *Lagrange's interpolation formula*

$$p(x) = y_1 L_1(x) + y_2 L_2(x) + \cdots + y_n L_n(x) \qquad (0.9.11.4)$$

for a polynomial $p(x)$ of degree at most $n-1$ that satisfies the equations (0.9.11.3).

0.10 Change of basis

Let V be an n-dimensional vector space over the field $\mathbf{F}$, and let $\mathcal{B}_1 = \{v_1, v_2, \ldots, v_n\}$ be a basis for V. If $x \in V$ is any given vector, then there exists some representation of $x = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n$ because the set $\mathcal{B}_1$ spans V. If there were some other representation of $x = \beta_1 v_1 + \beta_2 v_2 + \cdots + \beta_n v_n$ in the same basis, then

$$0 = x - x = (\alpha_1 - \beta_1)v_1 + (\alpha_2 - \beta_2)v_2 + \cdots + (\alpha_n - \beta_n)v_n$$

from which it follows that all $\alpha_i - \beta_i = 0$ because the set $\mathcal{B}_1$ is independent. Given the basis $\mathcal{B}_1$, the linear mapping

$$x \to [x]_{\mathcal{B}_1} \equiv \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix}, \qquad \text{where } x = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n$$

from V to $\mathbf{F}^n$ is well defined, one-to-one, and onto. The scalars α_i are called the *coordinates* of x with respect to the basis $\mathcal{B}_1$, and the column vector $[x]_{\mathcal{B}_1}$ is the unique $\mathcal{B}_1$ *coordinate representation of* x.

Let $T: V \to V$ be a given linear transformation. The action of T on any $x \in V$ is determined once one knows the n vectors $Tv_1, Tv_2, \ldots, Tv_n$, because any $x \in V$ has a unique representation $x = \alpha_1 v_1 + \cdots + \alpha_n v_n$ and $Tx = T(\alpha_1 v_1 + \cdots + \alpha_n v_n) = T(\alpha_1 v_1) + \cdots + T(\alpha_n v_n) = \alpha_1 Tv_1 + \cdots + \alpha_n Tv_n$ by linearity. Thus, the value of Tx is determined once $[x]_{\mathcal{B}_1}$ is known.

Let $\mathcal{B}_2 = \{w_1, w_2, \ldots, w_n\}$ be another, possibly different, basis for V, and suppose that the $\mathcal{B}_2$ coordinate representation of Tv_j is

$$[Tv_j]_{\mathcal{B}_2} = \begin{bmatrix} t_{1j} \\ t_{2j} \\ \vdots \\ t_{nj} \end{bmatrix}, \qquad j = 1, 2, \ldots, n$$

Then for any $x \in V$ we have

$$[Tx]_{\mathcal{B}_2} = \left[\sum_{j=1}^{n} \alpha_j T v_j\right]_{\mathcal{B}_2} = \sum_{j=1}^{n} \alpha_j [Tv_j]_{\mathcal{B}_2}$$

$$= \sum_{j=1}^{n} \alpha_j \begin{bmatrix} t_{1j} \\ t_{2j} \\ \vdots \\ t_{nj} \end{bmatrix} = \begin{bmatrix} t_{11} & \cdots & t_{1n} \\ \vdots & \ddots & \vdots \\ t_{n1} & \cdots & t_{nn} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}$$

The n-by-n array $[t_{ij}]$ depends on T and on the choice of the bases $\mathcal{B}_1$ and $\mathcal{B}_2$, but it does not depend on x. We define the $\mathcal{B}_1$-$\mathcal{B}_2$ *basis representation of* T to be

$$_{\mathcal{B}_2}[T]_{\mathcal{B}_1} = \begin{bmatrix} t_{11} & \cdots & t_{1n} \\ \vdots & \ddots & \vdots \\ t_{n1} & \cdots & t_{nn} \end{bmatrix} = [[Tv_1]_{\mathcal{B}_2} \cdots [Tv_n]_{\mathcal{B}_2}]$$

We have just shown that $[Tx]_{\mathcal{B}_2} = {}_{\mathcal{B}_2}[T]_{\mathcal{B}_1}[x]_{\mathcal{B}_1}$ for any $x \in V$. In practice, the case $\mathcal{B}_2 = \mathcal{B}_1$ is the most common one for presenting a basis representation of T; ${}_{\mathcal{B}_1}[T]_{\mathcal{B}_1}$ is called the $\mathcal{B}_1$ *basis representation of* T.

Consider the identity linear transformation $I: V \to V$ defined by $Ix = x$ for all x. Then

$$[x]_{\mathcal{B}_2} = [Ix]_{\mathcal{B}_2} = {}_{\mathcal{B}_2}[I]_{\mathcal{B}_1}[x]_{\mathcal{B}_1} = {}_{\mathcal{B}_2}[I]_{\mathcal{B}_1}[Ix]_{\mathcal{B}_1} = {}_{\mathcal{B}_2}[I]_{\mathcal{B}_1}\,{}_{\mathcal{B}_1}[I]_{\mathcal{B}_2}[x]_{\mathcal{B}_2}$$

for all $x \in V$. By successively choosing $x = w_1, w_2, \ldots, w_n$, this identity permits us to identify each column of ${}_{\mathcal{B}_2}[I]_{\mathcal{B}_1}\,{}_{\mathcal{B}_1}[I]_{\mathcal{B}_2}$ and shows that

$${}_{\mathcal{B}_2}[I]_{\mathcal{B}_1}\,{}_{\mathcal{B}_1}[I]_{\mathcal{B}_2} = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix} = I$$

We commit a common abuse of notation by using I to denote the n-by-n identity matrix as well as the identity linear transformation. If we do the same calculation starting with $[x]_{\mathcal{B}_1} = [Ix]_{\mathcal{B}_1} = \cdots$, we also find that

$${}_{\mathcal{B}_1}[I]_{\mathcal{B}_2}\,{}_{\mathcal{B}_2}[I]_{\mathcal{B}_1} = I$$

Thus, the matrix ${}_{\mathcal{B}_2}[I]_{\mathcal{B}_1}$ is the matrix inverse of the matrix ${}_{\mathcal{B}_1}[I]_{\mathcal{B}_2}$. If we write $S \equiv {}_{\mathcal{B}_2}[I]_{\mathcal{B}_1}$, then $S^{-1} = {}_{\mathcal{B}_1}[I]_{\mathcal{B}_2}$. Thus, every matrix of the form ${}_{\mathcal{B}_2}[I]_{\mathcal{B}_1}$ is invertible. Conversely, every invertible matrix $S = [s_1\, s_2 \cdots s_n] \in M_n(\mathbf{F})$ is of the form ${}_{\mathcal{B}_1}[I]_{\mathcal{B}}$ for some basis $\mathcal{B}$. We may take $\mathcal{B}$ to be the vectors $\{\tilde{s}_1, \tilde{s}_2, \ldots, \tilde{s}_n\}$ defined by $[\tilde{s}_i]_{\mathcal{B}_1} = s_i$, $i = 1, 2, \ldots, n$. The set $\mathcal{B}$ is independent because S is invertible.

Notice that

$$_{\mathcal{B}_2}[I]_{\mathcal{B}_1} = [[Iv_1]_{\mathcal{B}_2} \cdots [Iv_n]_{\mathcal{B}_2}] = [[v_1]_{\mathcal{B}_2} \cdots [v_n]_{\mathcal{B}_2}]$$

so $_{\mathcal{B}_2}[I]_{\mathcal{B}_1}$ expresses the elements of the basis $\mathcal{B}_1$ in terms of the basis $\mathcal{B}_2$. Now let $x \in V$ and compute

$$\begin{aligned} _{\mathcal{B}_2}[T]_{\mathcal{B}_2}[x]_{\mathcal{B}_2} &= [Tx]_{\mathcal{B}_2} = [I(Tx)]_{\mathcal{B}_2} = {}_{\mathcal{B}_2}[I]_{\mathcal{B}_1}[Tx]_{\mathcal{B}_1} \\ &= {}_{\mathcal{B}_2}[I]_{\mathcal{B}_1}\, {}_{\mathcal{B}_1}[T]_{\mathcal{B}_1}[x]_{\mathcal{B}_1} = {}_{\mathcal{B}_2}[I]_{\mathcal{B}_1}\, {}_{\mathcal{B}_1}[T]_{\mathcal{B}_1}[Ix]_{\mathcal{B}_1} \\ &= {}_{\mathcal{B}_2}[I]_{\mathcal{B}_1}\, {}_{\mathcal{B}_1}[T]_{\mathcal{B}_1}\, {}_{\mathcal{B}_1}[I]_{\mathcal{B}_2}[x]_{\mathcal{B}_2} \end{aligned}$$

By choosing $x = w_1, w_2, \ldots, w_n$ successively, we conclude that

$$_{\mathcal{B}_2}[T]_{\mathcal{B}_2} = {}_{\mathcal{B}_2}[I]_{\mathcal{B}_1}\, {}_{\mathcal{B}_1}[T]_{\mathcal{B}_1}\, {}_{\mathcal{B}_1}[I]_{\mathcal{B}_2}$$

This identity shows how the basis representation of T changes if the basis used to compute the representation is changed. For this reason, the matrix $_{\mathcal{B}_2}[I]_{\mathcal{B}_1}$ is called the $\mathcal{B}_1 - \mathcal{B}_2$ *change of basis matrix.*

Any matrix $A \in M_n(\mathbf{F})$ is a basis representation of some linear transformation $T: V \to V$, for if $\mathcal{B}$ is any basis of V, we can determine Tx by $[Tx]_{\mathcal{B}} = A[x]_{\mathcal{B}}$. One computes easily that, for this T, $_{\mathcal{B}}[T]_{\mathcal{B}} = A$.

CHAPTER 1

Eigenvalues, eigenvectors, and similarity

1.0 Introduction

In this and all the following chapters, we motivate some key issues discussed in the chapter with examples of how they arise, either conceptually or in application.

1.0.1 **Change of basis and similarity.** Every invertible matrix is a change-of-basis matrix, and every change-of-basis matrix is invertible [see Section (0.10)]. Thus, if $\mathcal{B}$ is a given basis of a vector space V, if T is a given linear transformation on V, and if $A = {}_{\mathcal{B}}[T]_{\mathcal{B}}$ is the $\mathcal{B}$ basis representation of T, the set of all possible basis representations of T is

$$\begin{aligned} &\{{}_{\mathcal{B}_1}[I]_{\mathcal{B}}\, {}_{\mathcal{B}}[T]_{\mathcal{B}}\, {}_{\mathcal{B}}[I]_{\mathcal{B}_1} : \mathcal{B}_1 \text{ is a basis of } V\} \\ &\quad = \{S^{-1}AS : S \in M_n(\mathbf{F}) \text{ is an invertible matrix}\} \end{aligned}$$

This is just the set of all matrices that are *similar* to the given matrix A. Similar but not identical matrices are therefore just different basis representations of a single linear transformation.

One would expect similar matrices to share many important properties – at least, those properties that are intrinsic to the underlying linear transformation – and this is an important theme in linear algebra. It is often useful to step back from a question about a matrix to a question about some intrinsic property of the linear transformation of which it is only one of many possible representations.

The notion of similarity is a key concept in this chapter.

1.0.2 **Constrained extrema and eigenvalues.** A second key concept in this chapter is the notion of eigenvector and eigenvalue. We shall see that nonzero vectors x such that Ax is a multiple of x play a major role in analyzing the structure of a general matrix or linear transformation, but such vectors arise in the more elementary context of maximizing (or minimizing) a real symmetric quadratic form subject to a geometric constraint:

$$\text{Maximize } x^TAx, \quad \text{subject to} \quad x \in \mathbf{R}^n, \quad x^Tx = 1$$

in which $A^T = A \in M_n(\mathbf{R})$ is given. A conventional approach to such a constrained optimization problem is to introduce the Lagrangian $L = x^TAx - \lambda x^Tx$. Necessary conditions for an extremum then are

$$0 = \nabla L = 2(Ax - \lambda x) = 0$$

Thus, if a vector $x \in \mathbf{R}^n$ with $x^Tx = 1$ (and hence $x \neq 0$) is to be an extremum of x^TAx, it must necessarily satisfy the equation $Ax = \lambda x$, and hence Ax is a multiple of x. Such a pair λ, x is called an *eigenvalue, eigenvector pair.*

Problems

1. Explain why the constrained extremum problem in (1.0.2) must have a solution, and conclude that every real symmetric matrix has at least one real eigenvalue. *Hint:* Apply Weierstrass's theorem (Appendix E) to the continuous function $f(x) = x^TAx$.

2. Let $A \in M_n(\mathbf{R})$ be symmetric ($A^T = A$). Show that the solution to

$$\text{Maximize } x^TAx, \quad \text{subject to} \quad x^Tx = 1$$

is the largest eigenvalue of A.

1.1 The eigenvalue–eigenvector equation

1.1.1 **Notation.** By $M_n(\mathbf{F})$ we denote the n-by-n matrices over a field $\mathbf{F}$, usually the real numbers $\mathbf{R}$ or the complex numbers $\mathbf{C}$. Most often, the facts discussed are valid in the setting of complex-entried matrices, in which case $M_n(\mathbf{C})$ is abbreviated as M_n. For the reader uninterested in the generality of complex matrices, it will seldom make a substantial difference in the exposition, the algebra, or the facts if the material is interpreted in terms of the real numbers. We caution, however, that there are major differences between $\mathbf{R}$ and $\mathbf{C}$, often having to do with roots of polynomials and other flexibility associated with the "larger" complex

field. Often, a real-entried matrix may best be thought of as a complex-entried matrix with restricted entries. Recall also that the set (vector space) of all real-entried (respectively complex-entried) n vectors is denoted by $\mathbf{R}^n$ (respectively $\mathbf{C}^n$), both interpreted as column vectors. Finally, the *transpose* (0.2.5) of $A = [a_{ij}] \in M_n(\mathbf{F})$ is the matrix $[a_{ji}] \in M_n(\mathbf{F})$, denoted by A^T, and, if $\mathbf{F} \subseteq \mathbf{C}$, the *Hermitian adjoint* is the conjugate transpose $[\bar{a}_{ji}]$ of A, denoted by A^*. Similarly, if $x \in \mathbf{F}^n$, x^T denotes the row vector with the same entries as x, and, if $\mathbf{F} \subseteq \mathbf{C}$, x^* denotes the row vector whose entries are the complex conjugates of those of x. Here the overbar $\bar{\cdot}$ denotes the complex conjugate (see Appendix A) of a complex scalar or the component-wise complex conjugate of a vector or matrix.

A matrix $A \in M_n$ is thought of as a linear transformation from $\mathbf{C}^n$ into $\mathbf{C}^n$ (with respect to some given basis of $\mathbf{C}^n$), but it is also useful to think of it as an array of numbers. It is the interplay between these two concepts of A, and what the array of numbers tells us about the linear transformation, that is the essence of matrix theory and a key to applications. Perhaps the single most important concept in matrix theory is the set of n numbers $\sigma(A)$ that we associate with A, its eigenvalues.

1.1.2 **Definition.** If $A \in M_n$ and $x \in \mathbf{C}^n$, we consider the equation

$$Ax = \lambda x, \quad x \neq 0 \tag{1.1.3}$$

where λ is a scalar. If a scalar λ and a nonzero vector x happen to satisfy this equation, then λ is called an *eigenvalue* of A and x is called an *eigenvector* of A associated with λ. Notice that the two occur inextricably as a pair, and that an eigenvector cannot be the zero vector.

1.1.4 **Definition.** The set of all $\lambda \in \mathbf{C}$ that are eigenvalues of $A \in M_n$ is called the *spectrum* of A and is denoted by $\sigma(A)$. The *spectral radius* of A is the nonnegative real number $\rho(A) = \max\{|\lambda| : \lambda \in \sigma(A)\}$. This is just the radius of the smallest disc centered at the origin in the complex plane that includes all the eigenvalues of A.

Exercise. If x is an eigenvector associated with the eigenvalue λ of A, show that any nonzero scalar multiple of x is an eigenvector also.

Even if they had no other importance, eigenvalues and eigenvectors would be interesting algebraically, since, according to (1.1.3), the eigenvectors are just those vectors such that multiplication by A has a very simple form – the same as multiplication by a scalar (the eigenvalue).

Example. Consider the matrix

$$A = \begin{bmatrix} 7 & -2 \\ 4 & 1 \end{bmatrix} \in M_2$$

Then we have $3 \in \sigma(A)$ with $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ as an associated eigenvector since

$$A\begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ 6 \end{bmatrix} = 3\begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Also, $5 \in \sigma(A)$. Find an eigenvector associated with the eigenvalue 5.

Recall that evaluation of a polynomial

$$p(t) = a_k t^k + a_{k-1} t^{k-1} + \cdots + a_1 t + a_0$$

at a matrix $A \in M_n$ is well defined since we may raise a square matrix to a positive integral power and may form linear combinations of matrices of the same size. Thus,

$$p(A) \equiv a_k A^k + a_{k-1} A^{k-1} + \cdots + a_1 A + a_0 I \tag{1.1.5}$$

It is useful to observe that a matrix related to $A \in M_n$ by the action of a polynomial has the same eigenvectors as A; its eigenvalues are linked to those of A in a simple way.

1.1.6 **Theorem.** Let $p(\cdot)$ be a given polynomial. If λ is an eigenvalue of $A \in M_n$, while x is an associated eigenvector, then $p(\lambda)$ is an eigenvalue of the matrix $p(A)$ and x is an eigenvector of $p(A)$ associated with $p(\lambda)$.

Proof: Consider $p(A)x$. First,

$$p(A)x \equiv a_k A^k x + a_{k-1} A^{k-1} x + \cdots + a_1 Ax + a_0 x$$

Second, $A^j x = A^{j-1} Ax = A^{j-1}\lambda x = \lambda A^{j-1} x = \cdots = \lambda^j x$ by repeated application of the eigenvalue-eigenvector equation. Thus,

$$p(A)x = a_k \lambda^k x + \cdots + a_0 x = (a_k \lambda^k + \cdots + a_0)x = p(\lambda)x. \quad \square$$

Exercise. If $\sigma(A) = \{-1, 2\}$, $A \in M_2$, what is $\sigma(A^2)$?

Exercise. If $D = \text{diag}(d_1, d_2, \ldots, d_n)$ is a diagonal matrix (0.9.1), what is $\sigma(D)$? Give an eigenvector associated with each eigenvalue. *Hint:* Consider the standard basis vectors e_i, $i = 1, \ldots, n$.

1.1.7 **Observation.** A matrix $A \in M_n$ is singular if and only if $0 \in \sigma(A)$.

Proof: The matrix A is singular if and only if $Ax = 0$ for some $x \neq 0$. This happens if and only if $Ax = 0x$ for some $x \neq 0$, that is, if and only if $\lambda = 0$ is an eigenvalue. $\square$

Problems

1. Suppose that $A \in M_n$ is nonsingular. According to (1.1.7), this is equivalent to saying that A has *no* eigenvalues equal to 0. If $\lambda \in \sigma(A)$, show that $\lambda^{-1} \in \sigma(A^{-1})$. If $Ax = \lambda x$ and $x \neq 0$, give an eigenvector of A^{-1} associated with λ^{-1}.

2. If the sum of the entries in each row of $A \in M_n$ is 1, show that $1 \in \sigma(A)$. *Hint:* Consider the vector $e = [1, 1, \ldots, 1]^T$ and observe that the row sums of A are all equal if and only if e is an eigenvector of A. If A is nonsingular, show that the row sums of A^{-1} are also 1. Given a polynomial $p(t)$, show that the row sums of $p(A)$ are all equal. To what?

3. Let $A \in M_n(\mathbf{R})$. If λ is a real eigenvalue of A with $Ax = \lambda x$, $0 \neq x \in \mathbf{C}^n$, let $x = \xi + i\eta$, where $\xi, \eta \in \mathbf{R}^n$ are the entrywise real and imaginary parts of x. Show that $A\xi = \lambda\xi$ and $A\eta = \lambda\eta$; conclude that there is a real eigenvector of A associated with λ. Must both ξ and η be eigenvectors of A? Can there be a real eigenvector associated with a complex non-real eigenvalue of A?

4. Consider the block diagonal matrix (0.9.2)

$$A = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}, \qquad A_{ii} \in M_{n_i}$$

Show that the eigenvalues of A are those of A_{11} together with those of A_{22}. *Hint:* First express the eigenvectors of A in terms of those of A_{11} and A_{22}.

5. $A \in M_n$ is called *idempotent* if $A^2 = A$. Show that each eigenvalue of an idempotent matrix is either 0 or 1.

6. $A \in M_n$ is called *nilpotent* if $A^q = 0$ for some positive integer q. The minimum such q is called the *index of nilpotence.* Show that all eigenvalues of a nilpotent matrix are 0. In the process, give an example of a nonzero matrix all of whose eigenvalues are equal to 0.

7. As we shall see, in the finite-dimensional setting upon which we concentrate, every complex or real square matrix has a complex eigenvalue.

However, it is possible for a linear transformation on an infinite dimensional vector space to have *no* eigenvalues whatsoever. Let V be the vector space of all formal infinite sequences of complex numbers:

$$V = \{(a_1, a_2, \dots, a_k, \dots) : a_i \in \mathbf{C}, \quad i = 1, 2, \dots\}$$

and define a linear transformation S on V by

$$S(a_1, a_2, \dots) = (0, a_1, a_2, \dots)$$

This transformation is sometimes called the *shift operator*. Verify that S is a linear transformation and show that S has no eigenvalues. *Hint:* Show that for a vector to be an eigenvector, all its components would have to be the same, but the only possible common value would be 0. The proposed vector would therefore have to be the zero vector, which cannot be an eigenvector.

8. A matrix $A \in M_n$ is called *Hermitian* if $A^* = A$ (0.2.5). If A is Hermitian, show that all eigenvalues of A are *real*. *Hint:* Let $\lambda \in \sigma(A)$ be arbitrary, and let x be an associated eigenvector. Then (1.1.3) implies that $x^*Ax = \lambda x^*x$. But $\overline{x^*Ax} = x^*A^*x = x^*Ax$, so that x^*Ax is real. Since x^*x is positive, $\lambda = x^*Ax/x^*x$ is real also.

1.2 The characteristic polynomial

Natural questions to ask about the eigenvalues of $A \in M_n$ are: How many are there? and, How may they be characterized?

The eigenvalue-eigenvector equation (1.1.3) may be rewritten equivalently as

$$(\lambda I - A)x = 0, \qquad x \neq 0 \tag{1.2.1}$$

Thus, $\lambda \in \sigma(A)$ if and only if $\lambda I - A$ is a singular matrix, that is,

$$\det(\lambda I - A) = 0 \tag{1.2.2}$$

1.2.3 **Definition.** Thought of as a formal polynomial in t, the *characteristic polynomial* of $A \in M_n$ is defined by

$$p_A(t) \equiv \det(tI - A)$$

Note: We use t as the formal variable in the characteristic polynomial to distinguish it from λ, a generic eigenvalue or zero of the polynomial. Elsewhere, the same symbol is sometimes used for both.

1.2.4 **Observation.** If $A \in M_n$, the characteristic polynomial $p_A(\cdot)$ has degree n and the set of roots of $p_A(t) = 0$ coincides with $\sigma(A)$.

Proof: That $p_A(\cdot)$ has degree n follows inductively from Laplace expansion of $\det(tI - A)$ by minors: each row of $tI - A$ contributes one and only one power of t as the determinant is expanded. The second statement is the equivalence of (1.1.3) and (1.2.2). □

Exercise. Show that the roots of $\det(A - tI) = 0$ are the same as those of $\det(tI - A) = 0$ and that $\det(A - tI) = (-1)^n \det(tI - A)$. Thus, the characteristic polynomial could alternatively be (and sometimes is) defined as $\det(A - tI)$. Show that the convention we have chosen insures that the (leading) coefficient of t^n is always $+1$.

Exercise. If $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$, show that $p_A(t) = t^2 - (a + d)t + (ad - bc)$ and that

$$\sigma(A) = \left\{ \frac{a + d \pm \sqrt{(a-d)^2 + 4bc}}{2} \right\}$$

If $A \in M_2(\mathbf{R})$, show that the eigenvalues of A are real if $bc \geq 0$. Furthermore, they are real if and only if $(a-d)^2 + 4bc \geq 0$. If they are not real, they occur as a complex conjugate pair. Finally, show that the eigenvalues are distinct if $(a-d)^2 + 4bc \neq 0$.

In certain general situations the eigenvalues of a matrix are easy to perceive. Most often, these are situations in which the determinant is easy to calculate due to the form of the matrix. These include diagonal and triangular matrices and some other special situations.

Exercise. Show that if $T \in M_n$ is triangular,

$$T = \begin{bmatrix} t_{11} & \cdots & t_{1n} \\ & \ddots & \vdots \\ 0 & & t_{nn} \end{bmatrix}$$

then $\sigma(T) = \{t_{11}, t_{22}, \dots, t_{nn}\}$, the diagonal entries of T.

Exercise. Each entry of the matrix $J_n \in M_n$ is equal to 1:

$$J_n = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \vdots & & \ddots & \vdots \\ 1 & & \cdots & 1 \end{bmatrix}$$

What are the eigenvalues of J_2? Show that 0 (occurring twice) and 3 are the only eigenvalues of J_3. What is the analog of this for general n? *Hint:* Consider the vector $e = [1, 1, \dots, 1]^T$.

Exercise. Determine all the eigenvalues, and an associated eigenvector for each, of the matrix

$$A=\begin{bmatrix}3 & -1 & -1\\ -1 & 3 & -1\\ -1 & -1 & 3\end{bmatrix}$$

Hint: Use the previous exercise, and write $A=4I-J_3$.

1.2.5 **Definition.** Recall from (0.7.1) that a k-by-k *principal submatrix* of $A\in M_n$ is one lying in the same set of k rows and columns and that a k-by-k *principal minor* is the determinant of such a principal submatrix. There are $\binom{n}{k}$ different k-by-k principal minors of $A=[a_{ij}]$, and the sum of these is denoted by $E_k(A)$. In particular, $E_1(A)=\sum_{i=1}^{n} a_{ii}$ is called the *trace* of A and is usually denoted by $\operatorname{tr} A$ or $\operatorname{trace} A$. Note that $E_n(A)=\det A$.

Exercise. If $A\in M_2$, show that $p_A(t)=t^2-(\operatorname{tr} A)t+\det A$ and that $\sum_{\lambda\in\sigma(A)}\lambda=\operatorname{tr} A$ and $\prod_{\lambda\in\sigma(A)}\lambda=\det A$.

A fundamental and nontrivial fact, called the *fundamental theorem of algebra* (Appendix C), is that a polynomial of degree n with complex coefficients has exactly n zeroes, counting multiplicities, among the complex numbers. In view of this we may make the very important following observation.

1.2.6 **Observation.** Each matrix $A\in M_n$ has, among the complex numbers, exactly n eigenvalues, counting multiplicities.

Note: At this point, when we refer to the "multiplicity" of an eigenvalue λ of $A\in M_n$, we simply mean the number of times λ occurs as a zero of the characteristic polynomial $p_A(\cdot)$. A more thorough discussion of multiplicity of eigenvalues will come in Section (1.4), but it is useful to know that there is a connection between the derivatives of a polynomial and the multiplicity of a zero of the polynomial. A polynomial $p(t)$ has λ as a zero of multiplicity $k\geq 1$ if and only if we can write $p(t)$ in the form $p(t)=(t-\lambda)^k q(t)$, where $q(t)$ is a polynomial such that $q(\lambda)\neq 0$. Differentiating this identity gives $p'(t)=k(t-\lambda)^{k-1}q(t)+(t-\lambda)^k q'(t)$, and from this representation it is clear that $p'(\lambda)=0$ if and only if $k>1$. If $k>1$, $p''(t)=k(k-1)(t-\lambda)^{k-2}q(t)+$polynomial terms each involving a factor $(t-\lambda)^m$ with $m\geq k-1$, so $p''(\lambda)=0$ if and only if $k>2$. Repetition of this calculation shows that λ is a zero of $p(t)$ of multiplicity k if and only if $p(\lambda)=p'(\lambda)=\cdots=p^{(k-1)}(\lambda)=0$ and $p^{(k)}(\lambda)\neq 0$.

1.2.7 **Examples.** The statement (1.2.6) depends heavily on the fact that the complex field is *algebraically closed,* that is, each polynomial of degree n with coefficients from the field has n zeroes in the field. For matrices over other fields, such as the real numbers or rational numbers, little in general can be said about how many eigenvalues a matrix has in the field. See Problem 8 of Section (1.1) for an example of something that can be said, however. Also, in the case of any field, a matrix could have very few different eigenvalues. The matrix

$$\begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix} \tag{1.2.7a}$$

has no real eigenvalues, though all its entries are real. The matrix

$$\begin{bmatrix}1 & 1 & & & & \\ & 1 & 1 & & 0 & \\ & & 1 & 1 & & \\ & & & \ddots & \ddots & \\ & 0 & & & & 1\\ & & & & & 1\end{bmatrix} \tag{1.2.7b}$$

has only one distinct eigenvalue (1 with multiplicity n), regardless of its size.

Exercise. Verify the statements in (1.2.7).

Exercise. If $A\in M_n(\mathbf{R})$ and if n is *odd*, show that A has at least one real eigenvalue. *Hint:* Recall that any nonreal complex zeroes of a polynomial with real coefficients must occur in conjugate pairs and note that $p_A(\cdot)$ has real coefficients if $A\in M_n(\mathbf{R})$.

In view of (1.2.6), we may list the eigenvalues of $A\in M_n$ as

$$\lambda_1,\lambda_2,\ldots,\lambda_n$$

where the ordering is arbitrary and we repeat eigenvalues according to multiplicity. Then, because of (1.2.4), we know that

$$p_A(t)=(t-\lambda_1)(t-\lambda_2)\cdots(t-\lambda_n) \tag{1.2.8}$$

1.2.9 **Definition.** The kth *elementary symmetric function* of the n numbers $\lambda_1,\ldots,\lambda_n$, $k\leq n$, is

$$S_k(\lambda_1,\ldots,\lambda_n)\equiv\sum_{1\leq i_1<\cdots<i_k\leq n}\prod_{j=1}^{k}\lambda_{i_j}$$

the sum of all $\binom{n}{k}$ k-fold products of distinct items from $\lambda_1, \ldots, \lambda_n$.

For example, $S_1(\lambda_1, \ldots, \lambda_n) = \lambda_1 + \cdots + \lambda_n$ is the sum of the λ_i and $S_n(\lambda_1, \ldots, \lambda_n) = \lambda_1 \cdots \lambda_n$ is the product of the λ_i. Because of (1.2.8) and the fact that $p_A(t)$ is defined by a certain determinant, there is a connection between the elementary symmetric functions $S_k(\lambda_1, \ldots, \lambda_n)$ of the eigenvalues of a matrix A and the $E_k(A)$, the sums of the k-by-k principal minors of A (1.2.5). The following two identities are straightforward, if laborious, to verify:

$$(t-\lambda_1)\cdots(t-\lambda_n) = t^n - S_1(\lambda_1, \ldots, \lambda_n)t^{n-1} + S_2(\lambda_1, \ldots, \lambda_n)t^{n-2} - \cdots \pm S_n(\lambda_1, \ldots, \lambda_n) \qquad (1.2.10)$$

and

$$p_A(t) = t^n - E_1(A)t^{n-1} + E_2(A)t^{n-2} - \cdots \pm E_n(A) \qquad (1.2.11)$$

Exercise. Convince yourself of (1.2.10) and (1.2.11). The former may be verified directly by picking out the coefficient of t^{n-k} in the product $(t-\lambda_1)\cdots(t-\lambda_n)$, and the latter may be verified inductively by Laplace expansion.

Combining (1.2.10) and (1.2.11) with (1.2.8), we have the following theorem.

1.2.12 **Theorem.** If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $A \in M_n$, then

$$S_k(\lambda_1, \ldots, \lambda_n) = E_k(A)$$

The kth elementary symmetric function of the eigenvalues of A is the sum of the k-by-k principal minors of A. In particular

$$\operatorname{tr} A = \sum_{i=1}^{n} \lambda_i$$

and

$$\det A = \prod_{i=1}^{n} \lambda_i$$

Problems

1. Verify (1.1.7) using (1.2.12).

2. For matrices $A \in M_{m,n}$ and $B \in M_{n,m}$ [see (0.2.1)], show by direct calculation that $\operatorname{tr} AB = \operatorname{tr} BA$. Use this fact to show that for $A \in M_n$ and nonsingular $S \in M_n$, $\operatorname{tr} S^{-1}AS = \operatorname{tr} A$. The matrix $S^{-1}AS$ is called a *similarity* of A, and this result says that the trace is a similarity invariant.

Similarity is the subject of the next section, and we shall find that all the principal minor sums $E_k(A)$ are similarity invariants. Note that the determinant is trivially a similarity invariant because of multiplicativity.

3. If $D \in M_n$ is diagonal, compute the characteristic polynomial $p_D(t)$ and show that $p_D(D) = 0$.

4. Let $A \in M_n$, and let $A_i = A(\{i\}') \in M_{n-1}$, the principal submatrix of A resulting from deleting row and column i, $i = 1, \ldots, n$. Show that

$$\frac{d}{dt} p_A(t) = \sum_{i=1}^{n} p_{A_i}(t) \qquad (1.2.13)$$

5. Recall Problem 6 of the preceding section. Show that the trace of a nilpotent matrix is 0. What is the characteristic polynomial of a nilpotent matrix?

6. If $\lambda \in \sigma(A)$ has multiplicity 1 as a root of $p_A(t) = 0$, $A \in M_n$, show that rank $(A - \lambda I) = n - 1$, but not necessarily conversely. [Recall example (1.2.7b).] *Hint:* Use (1.2.13) and the fact that $(d/dt)p_A(t) \neq 0$ at $t = \lambda$ to conclude that some principal submatrix of $A - \lambda I$ of size $n-1$ is nonsingular.

7. Use (1.2.12) to determine the characteristic polynomial of the matrix

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Consider how this procedure could be used to compute the characteristic polynomial of a general n-by-n tridiagonal matrix (0.9.10).

8. If $A \in M_n$ and $\sigma(A) = \{\lambda_1, \ldots, \lambda_n\}$, assume that $\sigma(A^k) = \{\lambda_1^k, \ldots, \lambda_n^k\}$. Show

$$\operatorname{tr} A^k = \sum_{i=1}^{n} \lambda_i^k$$

for all positive integers k. The right-hand sum is called the *kth moment* of the eigenvalues of A. The stated assumptions follow from (2.3.1).

9. Explicitly compute $S_2(\lambda_1, \ldots, \lambda_6)$, $S_3(\lambda_1, \ldots, \lambda_6)$, $S_4(\lambda_1, \ldots, \lambda_6)$, and $S_5(\lambda_1, \ldots, \lambda_6)$.

10. Let V be a vector space over a field **F**. An eigenvalue of a linear transformation $T: V \to V$ is a scalar $\lambda \in \mathbf{F}$ such that there is a nonzero vector $v \in V$ with $Tv = \lambda v$. Show that if **F** is the field of complex numbers

and if V is finite-dimensional, then every linear transformation T has an eigenvalue. Give examples to show that if either hypothesis is weakened (finite dimensionality of V or $\mathbf{F}=\mathbf{C}$), then T may not have an eigenvalue. *Hint:* Let $\mathcal{B}$ be a basis for V and consider $[T]_{\mathcal{B}}$.

11. Let $p(t)=a_nt^n+a_{n-1}t^{n-1}+\cdots+a_1t+a_0$, $a_n=1$, be a given monic polynomial with zeroes $\lambda_1,\lambda_2,\ldots,\lambda_n$, including multiplicities. Denote the kth moments of the zeroes by $\mu_k=\lambda_1^k+\lambda_2^k+\cdots+\lambda_n^k$, $k=1,2,\ldots$. Demonstrate *Newton's identities*

$$ka_{n-k}+\mu_1a_{n-k+1}+\mu_2a_{n-k+2}+\cdots+\mu_ka_n=0, \qquad k=1,2,\ldots,n \quad (1.2.14)$$

Explain why the first n moments of the zeroes uniquely determine the coefficients of the polynomial $p(t)$ (and hence the zeroes) and conversely. *Hint:* Show that for some $R>0$, if $|t|>R$, then $(t-\lambda_i)^{-1}=t^{-1}+\lambda_it^{-2}+\lambda_i^2t^{-3}+\cdots$ and hence

$$f(t)\equiv\sum_{i=1}^{n}(t-\lambda_i)^{-1}=nt^{-1}+\mu_1t^{-2}+\mu_2t^{-3}+\cdots \quad \text{for} \quad |t|>R$$

Show that $p'(t)=p(t)f(t)$, from which the Newton identities and the additional identities

$$\mu_ka_0+\mu_{k+1}a_1+\cdots+\mu_{n+k-1}a_{n-1}+\mu_{n+k}a_n=0, \qquad k=1,2,\ldots$$

for the higher-order moments follow from a comparison of coefficients.

12. Let $A,B\in M_n$ be given. Show that A and B have the same eigenvalues if and only if $\operatorname{tr}A^k=\operatorname{tr}B^k$ for $k=1,2,\ldots,n$. *Hint:* Use Problem 8 and the Newton identities (1.2.14) to show that the characteristic polynomials of A and B are the same.

1.3 Similarity

As indicated in Section (1.0), a similarity transformation of a matrix in M_n corresponds to representation of a linear transformation on $\mathbf{C}^n$ in another basis. Thus, studying similarity can be thought of as studying properties which are intrinsic to a linear transformation, or the properties which are common to all its various basis representations.

1.3.1 **Definition.** A matrix $B\in M_n$ is said to be *similar* to a matrix $A\in M_n$ if there exists a nonsingular matrix $S\in M_n$ such that

$$B=S^{-1}AS$$

The transformation $A\to S^{-1}AS$ is called a *similarity transformation* by the *similarity matrix S*. The relation "B is similar to A" is sometimes abbreviated $B\sim A$.

1.3.2 **Observation.** Similarity is an *equivalence relation* on M_n; that is, similarity is

(a) reflexive: $A\sim A$;
(b) symmetric: $B\sim A$ implies $A\sim B$; and
(c) transitive: $C\sim B$ and $B\sim A$ imply $C\sim A$.

Exercise. Verify (1.3.2).

Like any equivalence relation, the similarity relation partitions the set M_n into disjoint equivalence classes. Each equivalence class is the set of all matrices in M_n similar to a given matrix, a representative of the class. All matrices in an equivalence class are similar, and matrices in two different classes are not. Because of transitivity, the first and last matrices in an arbitrary finite sequence of similar matrices are in the same similarity equivalence class. The crucial observation is that matrices in an equivalence class share many important properties. Some of these will be mentioned here, but a more complete description of the *similarity invariants* (e.g., Jordan canonical form) will come later in Chapter 3.

1.3.3 **Theorem.** Let $A,B\in M_n$. If B is similar to A, then the characteristic polynomial of B is the same as that of A.

Proof: For any t we have

$$\begin{aligned}p_B(t)&=\det(tI-B)\\&=\det(tS^{-1}S-S^{-1}AS)=\det S^{-1}(tI-A)S\\&=\det S^{-1}\det(tI-A)\det S=(\det S)^{-1}(\det S)\det(tI-A)\\&=\det(tI-A)=p_A(t)\end{aligned}$$

□

1.3.4 **Corollary.** If $A,B\in M_n$ and if A and B are similar, then they have the same eigenvalues, counting multiplicity.

1.3.5 **Example.** Having the same eigenvalues is a necessary but not sufficient condition for similarity. Consider the matrices

$$\begin{bmatrix}0&1\\0&0\end{bmatrix} \quad \text{and} \quad \begin{bmatrix}0&0\\0&0\end{bmatrix}$$

Each has the eigenvalue 0 with multiplicity 2, but they are not similar.

Exercise. Show that the only matrix similar to the zero matrix is itself and use this fact to verify the statements in Example (1.3.5).

Exercise. If the matrices $A, B \in M_n$ are similar and if $q(\cdot)$ is a polynomial, show that $q(A)$ and $q(B)$ are similar. In particular, show that $A+\alpha I$ and $B+\alpha I$ are similar if α is a scalar.

Exercise. If $A, B, C, D \in M_n$ and $A \sim B$ via the similarity matrix S and $C \sim D$ via the same similarity matrix S, show that $A+C \sim B+D$.

Exercise. If $A, S \in M_n$ and if S is nonsingular, show that $E_k(S^{-1}AS) = E_k(A)$ and, in particular, that $\det S^{-1}AS = \det A$ and $\operatorname{tr} S^{-1}AS = \operatorname{tr} A$, that is, the determinant, trace, and other sums of k-by-k principal minors are similarity invariants.

Exercise. Show that rank is also a similarity invariant: If $B \in M_n$ is similar to $A \in M_n$, then $\operatorname{rank} B = \operatorname{rank} A$. *Hint:* See (0.4.6).

Since diagonal matrices are especially simple and have very nice properties, it is of interest to know for which $A \in M_n$ there is a diagonal matrix in the similarity equivalence class of A, that is, which matrices are similar to diagonal matrices.

1.3.6 **Definition.** If the matrix $A \in M_n$ is similar to a diagonal matrix, then A is said to be *diagonalizable*. Sometimes the term *diagonable* is used.

1.3.7 **Theorem.** Let $A \in M_n$. Then A is diagonalizable if and only if there is a set of n linearly independent vectors, each of which is an eigenvector of A.

Proof: If A has n linearly independent eigenvectors $x^{(1)}, \ldots, x^{(n)}$, form a nonsingular matrix S with them as columns and calculate

$$\begin{aligned} S^{-1}AS &= S^{-1}[Ax^{(1)} Ax^{(2)} \ldots Ax^{(n)}] \\ &= S^{-1}[\lambda_1 x^{(1)} \ldots \lambda_n x^{(n)}] = S^{-1}[x^{(1)} \ldots x^{(n)}]\Lambda \\ &= S^{-1}S\Lambda = \Lambda \end{aligned}$$

where

$$\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$$

and $\lambda_1, \ldots, \lambda_n$ are eigenvalues of A.

Conversely, suppose that there is a similarity matrix S such that $S^{-1}AS = \Lambda$ is diagonal. Then $AS = S\Lambda$. This means that A times the ith column of S (i.e., the ith column of AS) is the ith diagonal entry of Λ times the ith column of S (i.e., the ith column of $S\Lambda$), or that the ith column of S is an eigenvector of A associated with the ith diagonal entry of Λ. Since S is nonsingular, there are n linearly independent eigenvectors. □

Note that the proof of (1.3.7) is, in principle, an algorithm for diagonalizing a diagonalizable matrix: find the eigenvalues of A; find associated eigenvectors, counting multiplicity, and array them in the matrix S. *If* the eigenvectors are linearly independent, then S is a diagonalizing similarity matrix. We stress, however, that, except for small analytical examples, this is *not* a practical computational procedure.

Remark: If $A \in M_n$ is diagonalizable, the diagonal entries of any diagonal matrix to which it is similar must be the eigenvalues of A, with proper multiplicities. Moreover, the linearly independent eigenvectors (which make up the similarity matrix) must correspond to the different eigenvalues with proper multiplicities; that is, if $x^{(1)}, \ldots, x^{(n)}$ are linearly independent eigenvectors and $p_A(t) = (t-\lambda_1)\cdots(t-\lambda_n)$, then $Ax^{(i)} = \lambda_{\tau(i)}x^{(i)}$ for some permutation τ of the indices.

Exercise. Show that the matrix $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ is *not* diagonalizable.

Reason, on the one hand, that if it were diagonalizable, it would be similar to the 0 matrix - which it is not; and, on the other hand, calculate that, up to a factor of scale, there is only one eigenvector associated with 0.

Exercise. If A is diagonalizable and if $q(\cdot)$ is a polynomial, show that $q(A)$ is diagonalizable. *Hint:* $q(SAS^{-1}) = Sq(A)S^{-1}$.

Exercise. If $A \in M_n$ and if $\lambda \in \sigma(A)$ has multiplicity m as an eigenvalue of A, show that A is not diagonalizable if $\operatorname{rank}(A-\lambda I) > n-m$.

A simple circumstance in which diagonalizability is assured is that in which the eigenvalues are distinct. An important precursor to this fact, which is otherwise useful, is the following lemma.

1.3.8 **Lemma.** Suppose that $\lambda_1, \ldots, \lambda_k$ are eigenvalues of $A \in M_n$, no two of which are the same, and suppose that $x^{(i)}$ is an eigenvector associated with λ_i, $i = 1, \ldots, k$. Then $\{x^{(1)}, \ldots, x^{(k)}\}$ is a linearly independent set.

Proof: The proof is essentially by contradiction. Suppose that $x^{(1)}, \ldots, x^{(k)}$ is actually a linearly dependent set. Then there is a nontrivial linear combination which produces the 0 vector, and in fact there is such a linear

combination with the *fewest* nonzero coefficients. Suppose that such a minimal linear dependence relation is

$$\alpha_1 x^{(1)} + \alpha_2 x^{(2)} + \cdots + \alpha_r x^{(r)} = 0, \qquad r \le k$$

We have $r > 1$ because all $x^{(i)} \neq 0$. We may assume for convenience (renumber if necessary) that it involves the first r vectors. We also have

$$A(\alpha_1 x^{(1)} + \cdots + \alpha_r x^{(r)}) = \alpha_1 A x^{(1)} + \cdots + \alpha_r A x^{(r)}$$
$$= \alpha_1 \lambda_1 x^{(1)} + \cdots + \alpha_r \lambda_r x^{(r)} = 0$$

another dependence relation. Now multiply the first dependence relation by λ_r and subtract it from the second to produce

$$\alpha_1(\lambda_1 - \lambda_r)x^{(1)} + \cdots + \alpha_{r-1}(\lambda_{r-1} - \lambda_r)x^{(r-1)} = 0$$

a third dependence relation, which has fewer nonzero coefficients than the first. This last relation is nontrivial since $\lambda_i \neq \lambda_r$, $i = 1, \ldots, r-1$. This contradicts the minimality assumption for the first dependence relation and completes the proof. □

1.3.9 **Theorem.** If $A \in M_n$ has n distinct eigenvalues, then A is diagonalizable.

Proof: If $\sigma(A) = \{\lambda_1, \ldots, \lambda_n\}$, let $x^{(i)}$ be an eigenvector associated with λ_i, $i = 1, \ldots, n$. Since the eigenvalues are all different, $\{x^{(1)}, \ldots, x^{(n)}\}$ is a linearly independent set by (1.3.8), and therefore A is diagonalizable by (1.3.7). □

Exercise. Give an example of a diagonalizable matrix $A \in M_n$ that does not have distinct eigenvalues.

Exercise. Recall from (0.9.5) that a permutation matrix P is a 0, 1 matrix with exactly one 1 in each row and column. Thus $P^T = P^{-1}$. Show that a permutation similarity of $A \in M_n$ reorders the diagonal entries of A, and show that for any diagonal matrix there is a permutation similarity with the diagonal entries occurring in any order, in particular, with any repeated diagonal entries occurring contiguously.

In general, matrices $A, B \in M_n$ do *not* commute under multiplication, but if A and B are both diagonal, they always commute. The latter observation can be generalized somewhat; the following lemma will be helpful in this regard.

1.3.10 **Lemma.** Let $A \in M_n$ and $B \in M_m$ be given matrices and let

$$C = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$$

be the direct sum of A and B. Then C is diagonalizable if and only if both A and B are diagonalizable.

Proof: If there is a nonsingular matrix $S_1 \in M_n$ such that $S_1^{-1}AS_1$ is diagonal and a nonsingular matrix $S_2 \in M_m$ such that $S_2^{-1}BS_2$ is diagonal, then one checks easily that $S^{-1}CS$ is diagonal if S is the direct sum

$$S \equiv \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix}$$

Conversely, if C is diagonalizable, there is a nonsingular matrix $S \in M_{n+m}$ such that $S^{-1}CS = \Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{n+m})$ is diagonal. If we write $S = [s_1\, s_2 \ldots s_{n+m}]$ with

$$s_i = \begin{bmatrix} \xi_i \\ \eta_i \end{bmatrix} \in \mathbf{C}^{n+m}, \qquad \xi_i \in \mathbf{C}^n, \qquad \eta_i \in \mathbf{C}^m \quad \text{for } i = 1, 2, \ldots, n+m$$

then $Cs_i = \lambda_i s_i$ implies that $A\xi_i = \lambda_i \xi_i$ and $B\eta_i = \lambda_i \eta_i$ for $i = 1, 2, \ldots, n+m$. If there were fewer than n independent vectors in the set $\{\xi_1, \ldots, \xi_{n+m}\}$, then the column rank (and hence the row rank) of the matrix

$$[\xi_1\, \xi_2 \ldots \xi_{n+m}] \in M_{n,n+m}$$

would be less than n. By the same reasoning, if there were fewer than m independent vectors in the set $\{\eta_1, \ldots, \eta_{n+m}\}$, then the column rank (and hence the row rank) of the matrix

$$[\eta_1\, \eta_2 \ldots \eta_{n+m}] \in M_{m,n+m}$$

would be less than m. In either event (or both), the matrix

$$S = [s_1 \ldots s_{m+n}] = \begin{bmatrix} \xi_1 \ldots \xi_{n+m} \\ \eta_1 \ldots \eta_{n+m} \end{bmatrix} \in M_{n+m}$$

would have row rank (and hence rank) less than $n+m$, which is impossible since S is invertible. Thus, there are exactly n independent vectors in the set $\{\xi_1, \xi_2, \ldots, \xi_{n+m}\}$, and since each is an eigenvector of A, the matrix A must be diagonalizable. The same argument shows that the matrix B is diagonalizable. □

1.3.11 **Definition.** Two diagonalizable matrices $A, B \in M_n$ are said to be *simultaneously diagonalizable* if there is a single similarity matrix $S \in M_n$

such that $S^{-1}AS$ and $S^{-1}BS$ are both diagonal, that is, if there is a single basis in which the representations of both linear transformations are diagonal.

Exercise. Show that if $A, B \in M_n$ are simultaneously diagonalizable, then they commute. *Hint:* Write $A = SDS^{-1}$ and $B = SES^{-1}$, D and E diagonal. Then calculate AB and BA, using the fact that diagonal matrices commute. This type of manipulation is used frequently.

Exercise. Show that if $A \in M_n$ is diagonalizable and if λI is a scalar matrix in M_n, then A and λI are simultaneously diagonalizable.

1.3.12 **Theorem.** Let $A, B \in M_n$ be diagonalizable. Then A and B commute if and only if they are simultaneously diagonalizable.

Proof: Assume that A and B commute, perform a similarity transformation on both A and B that diagonalizes A, and then assume without loss of generality that A is diagonal. Assume further, without loss of generality, that any multiple eigenvalues of A occur contiguously on the main diagonal. Since $AB = BA$ (the common similarity has not changed this), we have

$$\lambda_i b_{ij} = b_{ij} \lambda_j$$

where $B = [b_{ij}]$ and $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A. Since $(\lambda_i - \lambda_j) b_{ij} = 0$, we conclude that $b_{ij} = 0$ whenever $\lambda_i \neq \lambda_j$. Thus, given the ordering of the λ_i terms, B is a block diagonal matrix:

$$B = \begin{bmatrix} B_1 & & 0 \\ & \ddots & \\ 0 & & B_k \end{bmatrix} \tag{1.3.13}$$

where these is one block B_i for each different eigenvalue of A. Each B_i is square and has size equal to the multiplicity of the eigenvalue of A to which it corresponds. Since B is diagonalizable, each B_i is diagonalizable by (1.3.10). Let T_i be a nonsingular matrix such that $T_i^{-1} B_i T_i$ is diagonal. Since A has the partitioned form

$$A = \begin{bmatrix} \lambda_1 I & & & 0 \\ & \lambda_2 I & & \\ & & \ddots & \\ 0 & & & \lambda_k I \end{bmatrix} \tag{1.3.14}$$

where each scalar matrix $\lambda_i I$ is the same size as B_i, we see that $T^{-1}AT$ and $T^{-1}BT$ are both diagonal, where T is the direct sum

$$T = \begin{bmatrix} T_1 & & & 0 \\ & T_2 & & \\ & & \ddots & \\ 0 & & & T_k \end{bmatrix} \tag{1.3.15}$$

Note that $T_i^{-1} \lambda_i I T_i = \lambda_i I$.

The converse is included in an earlier exercise. □

We conclude this section by extending (1.3.12) to larger sets of matrices and by making a weaker statement in the case of nondiagonalizable matrices.

1.3.16 **Definitions.** A *family* $\mathfrak{F} \subseteq M_n$ of matrices is an arbitrary (finite or infinite) set of matrices, and a *commuting family* is one in which each pair in the set commutes under multiplication. A subspace $W \subseteq \mathbf{C}^n$ is said to be *A-invariant*, for $A \in M_n$, if $Aw \in W$ for every $w \in W$, and W is called *$\mathfrak{F}$-invariant*, for a family $\mathfrak{F} \subseteq M_n$, if W is A-invariant for each $A \in \mathfrak{F}$.

Notice that if $A \in M_n$, each nonzero element of a one-dimensional A-invariant subspace of $\mathbf{C}^n$ is an eigenvector of A.

Exercise. Let $A \in M_n$. If W is an A-invariant subspace of $\mathbf{C}^n$ of dimension at least 1, show that there is an eigenvector of A in W. *Hint:* Choose a basis for W and consider the matrix that is the basis representation of the linear transformation $T: w \to Aw$ on W. Argue that this matrix has an eigenvalue. Crucial point: Why is T a linear transformation on W?

A key observation is the following lemma.

1.3.17 **Lemma.** If $\mathfrak{F} \subseteq M_n$ is a commuting family, then there is a vector $x \in \mathbf{C}^n$ that is an eigenvector of *every* $A \in \mathfrak{F}$.

Proof: Let $W \subseteq \mathbf{C}^n$ be an $\mathfrak{F}$-invariant subspace of *minimum positive* dimension; such a W exists, but it need not be unique. Since $\mathbf{C}^n$ is itself $\mathfrak{F}$-invariant, we know there is an $\mathfrak{F}$-invariant subspace of dimension n. If there is one of dimension $n-1$, then ask if there is one of dimension $n-2$, and so forth. We actually show that every nonzero vector in W is an eigenvector of every $A \in \mathfrak{F}$, which is sufficient to complete the proof.

If this is not the case, then for some matrix $A \in \mathfrak{F}$, not every nonzero vector in W is an eigenvector of A. But, since W is $\mathfrak{F}$-invariant, it is A-invariant, and there is an $x \neq 0$ in W such that $Ax = \lambda x$ for some eigenvalue λ. Define $W_0 = \{y \in W: Ay = \lambda y\}$, so that $x \in W_0$ and $W_0 \subseteq W$ is a subspace. Because of the assumption about A, $W_0 \neq W$, so that the (positive) dimension of W_0 is strictly smaller than that of W. If $B \in \mathfrak{F}$, we have $Bx \in W$ if $x \in W_0$, since $W_0 \subseteq W$ and W is $\mathfrak{F}$-invariant. But then, since $\mathfrak{F}$ is a commuting family, $A(Bx) = (AB)x = (BA)x = B(Ax) = B\lambda x = \lambda(Bx)$ and we conclude that $Bx \in W_0$. It follows that W_0 is $\mathfrak{F}$-invariant. But since W_0 has strictly lower positive dimension than W, we reach a contradiction, which completes the proof. □

Lemma (1.3.17) concerns commuting families of arbitrary cardinality. In particular, if $\mathfrak{F} = \{A, B\}$ is a family of only two matrices, it says that any pair of commuting matrices has a common eigenvector. Theorem (1.3.12) says that if A and B not only commute but are each diagonalizable as well, then they are simultaneously diagonalizable. Our next result shows that there is nothing special about commuting families of two diagonalizable matrices; this result generalizes to families of arbitrary cardinality.

1.3.18 **Definition.** A *simultaneously diagonalizable* family $\mathfrak{F} \subset M_n$ is a family for which there is a single nonsingular matrix $S \in M_n$ such that $S^{-1}AS$ is diagonal for every $A \in \mathfrak{F}$.

1.3.19 **Theorem.** Let $\mathfrak{F} \subset M_n$ be a family of diagonalizable matrices. Then $\mathfrak{F}$ is a commuting family if and only if it is a simultaneously diagonalizable family.

Proof: If $\mathfrak{F}$ is simultaneously diagonalizable, then it is a commuting family by a previous exercise. We prove the converse by induction on n. If $n = 1$, there is nothing to prove since every family is both commuting and diagonal. Let us suppose that $n \geq 2$ and that, for $k = 1, 2, \ldots, n-1$, the assertion has been proved for all families of k-by-k matrices that satisfy the hypotheses. If every matrix in $\mathfrak{F}$ is a scalar matrix, there is nothing to prove, so we may assume that $A \in \mathfrak{F}$ is a given n-by-n diagonalizable matrix with at least two distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$, $2 \leq k \leq n$, that $AB = BA$ for every matrix $B \in \mathfrak{F}$, and that each $B \in \mathfrak{F}$ is diagonalizable. Using the same argument as in (1.3.12), we can reduce to the case in which A is actually a diagonal matrix, any multiple eigenvalues of A occur contiguously, and the ordering of the eigenvalues is fixed, that is, A has the form (1.3.14). Since every $B \in \mathfrak{F}$ commutes with A, the argument in (1.3.12) shows that every $B \in \mathfrak{F}$ has the form of a direct sum (1.3.13) of matrices, each of size $n-1$ or less. The sizes and locations of the blocks in (1.3.13) are determined solely by the multiplicities and ordering of the eigenvalues of A and are therefore the same for all $B \in \mathfrak{F}$. Since all the matrices $B \in \mathfrak{F}$ commute (not just with A), and since every $B \in \mathfrak{F}$ has the form of a direct sum (1.3.13), each of the k direct summand blocks of any matrix in $\mathfrak{F}$ must commute with the corresponding block of every other matrix in $\mathfrak{F}$, and each of these blocks is diagonalizable by (1.3.10). By the induction hypothesis, there are k similarity matrices $T_1, T_2, \ldots, T_k$ of appropriate size, each of which diagonalizes the corresponding block of every matrix in $\mathfrak{F}$. As in (1.3.15), the direct sum $T_1 \oplus \cdots \oplus T_k$ diagonalizes every matrix in $\mathfrak{F}$. □

Remarks: Two important issues related to this section will be deferred until Chapter 3: (1) Given $A, B \in M_n$, how can we determine if A is similar to B? This is a motivation for canonical forms under similarity. (2) How can we tell if a given matrix $A \in M_n$ is diagonalizable without computing its eigenvectors?

As a final remark on commutativity, we observe that although AB and BA are not necessarily the same matrix (and not even necessarily the same size even when both are defined), they are as much the same as possible from the point of view of their eigenvalues. If A and B are both square, AB and BA have exactly the same eigenvalues.

1.3.20 **Theorem.** Suppose that $A \in M_{m,n}$ and $B \in M_{n,m}$ with $m \leq n$. Then BA has the same eigenvalues as AB, counting multiplicity, together with an additional $n-m$ eigenvalues equal to 0; that is, $p_{BA}(t) = t^{n-m}p_{AB}(t)$. If $m = n$ and at least one of A or B is nonsingular, then AB and BA are similar.

Proof: Consider the following two identities involving block matrices in M_{m+n}:

$$\begin{bmatrix} AB & 0 \\ B & 0 \end{bmatrix}\begin{bmatrix} I & A \\ 0 & I \end{bmatrix} = \begin{bmatrix} AB & ABA \\ B & BA \end{bmatrix}$$

$$\begin{bmatrix} I & A \\ 0 & I \end{bmatrix}\begin{bmatrix} 0 & 0 \\ B & BA \end{bmatrix} = \begin{bmatrix} AB & ABA \\ B & BA \end{bmatrix}$$

Since the block matrix

$$\begin{bmatrix} I & A \\ 0 & I \end{bmatrix} \in M_{m+n}$$

is nonsingular (all its eigenvalues are +1), we conclude that

$$\begin{bmatrix} I & A \\ 0 & I \end{bmatrix}^{-1} \begin{bmatrix} AB & 0 \\ B & 0 \end{bmatrix} \begin{bmatrix} I & A \\ 0 & I \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ B & BA \end{bmatrix}$$

that is, the two $(m+n)$-by-$(m+n)$ matrices

$$C_1 = \begin{bmatrix} AB & 0 \\ B & 0 \end{bmatrix} \quad \text{and} \quad C_2 = \begin{bmatrix} 0 & 0 \\ B & BA \end{bmatrix}$$

are similar. The eigenvalues of C_1 are the eigenvalues of AB together with n zeroes. The eigenvalues of C_2 are the eigenvalues of BA together with m zeroes. Since the eigenvalues of C_1 and C_2 are the same by (1.3.4), including multiplicities, the main assertion of the theorem follows. The final assertion follows from the observation that $AB = A(BA)A^{-1}$ if A is nonsingular and $m = n$. □

Problems

1. If $A, B \in M_n$ and if A and B commute, show that A and any polynomial in B commute.

2. Let $A, B \in M_n$ with $\sigma(A) = \{\lambda_1, \ldots, \lambda_n\}$ and $\sigma(B) = \{\mu_1, \ldots, \mu_n\}$. If A and B are diagonalizable and commute, show that the eigenvalues of $A+B$ are

$$\lambda_1 + \mu_{i_1}, \lambda_2 + \mu_{i_2}, \ldots, \lambda_n + \mu_{i_n}$$

for some permutation $i_1, \ldots, i_n$ of $1, \ldots, n$.

3. If $A \in M_n$ and $A = S^{-1}DS$, $D = \text{diag}(d_1, \ldots, d_n)$, and $p(\bullet)$ is a polynomial, show that $p(A) = S^{-1}p(D)S$ and that $p(D) = \text{diag}(p(d_1), \ldots, p(d_n))$. This provides a simple way of evaluating $p(A)$ if one can diagonalize A.

4. Give an example of two commuting matrices that are not simultaneously diagonalizable. Does this contradict Theorem (1.3.12)?

5. If $A \in M_n$ has distinct eigenvalues and if A commutes with a given matrix $B \in M_n$, show that B is a polynomial in A of degree at most $n-1$. *Hint:* Use the method employed in the proof of Theorem (1.3.12) to show that B and A must be simultaneously diagonalizable. Then recall that, given distinct numbers $\alpha_1, \ldots, \alpha_n$ and numbers $\beta_1, \ldots, \beta_n$, there is a (Lagrange interpolating) polynomial $p(\bullet)$ of degree at most $n-1$ such that $p(\alpha_i) = \beta_i$. See (0.9.11).

6. If $A \in M_n$ is diagonalizable, consider the characteristic polynomial $p_A(t)$ and show that $p_A(A)$ is the zero matrix.

7. A matrix $A \in M_n$ is a *square root* of $B \in M_n$ if $A^2 = B$. Show that every diagonalizable matrix in M_n has a square root.

8. If $A, B \in M_n$ and if at least one has distinct eigenvalues (no assumption, even of diagonalizability, about the other), show that A and B commute if and only if they are simultaneously diagonalizable. *Suggestion:* One direction is easy; for the other, try the following type of argument as an alternate to that used for (1.3.12). Suppose B has distinct eigenvalues, $\lambda \in \sigma(B)$, and $Bx = \lambda x$ with $x \neq 0$. Then $B(Ax) = A(Bx) = A\lambda x = \lambda Ax$, which implies that Ax is also an eigenvector of B associated with λ. Since there cannot be two linearly independent such vectors (because λ has multiplicity 1), Ax must be a multiple μ of x; that is, $Ax = \mu x$. Thus, every eigenvector of B is also an eigenvector of A, and A is diagonalizable by the same matrix of eigenvectors that diagonalizes B. See Problems 12 and 13 for another approach to the same fact.

9. Provide the details for the following alternate proof of Theorem (1.3.20). (a) First, suppose that $A, B \in M_n$ and that at least one of them is nonsingular. Show that AB is similar to BA, and hence the characteristic polynomials of AB and BA are the same. *Hint:* If A is nonsingular, $BA = A^{-1}(AB)A$. Hence $\sigma(AB) = \sigma(BA)$ in this case. (b) Consider the singular matrices $A = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$ and $B = \left[\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right]$. Show that AB and BA are *not* similar, but that they do have the same eigenvalues. (c) Show that if $A, B \in M_n$, AB and BA have the same eigenvalues, counting multiplicities. *Hint:* Consider the following analytic argument. For all sufficiently small $\epsilon > 0$, $A_\epsilon \equiv A + \epsilon I$ is nonsingular; thus, $A_\epsilon B$ and BA_ϵ are similar and hence the characteristic polynomials of $A_\epsilon B$ and BA_ϵ are the same. If we now let $\epsilon \to 0$, similarity may fail in the limit, but equality of the characteristic polynomials continues to hold since $p_{A_\epsilon B}(t) = \det(tI - A_\epsilon B)$ depends continuously on ϵ. Thus, AB and BA have the same characteristic polynomials and therefore the *same eigenvalues,* counting multiplicities. (d) Finally, if $A \in M_{m,n}$ and $B \in M_{n,m}$, show that AB and BA have the same eigenvalues, counting multiplicities, *except* that BA has an additional $n-m$ eigenvalues equal to 0 (assuming $n > m$); equivalently, $p_{BA}(t) = t^{n-m}p_{AB}(t)$. *Hint:* Make n-by-n matrices out of both A (by appending 0 rows) and B (by appending 0 columns), apply the last result, and compare the two new products (appropriately partitioned) to the two old ones.

10. Use (1.3.8) to prove the following generalization: Let $A \in M_n$ be given, and let $\lambda_1, \ldots, \lambda_k$ be distinct eigenvalues of A. For each $i = 1, 2, \ldots, k$ suppose $\{x_1^{(i)}, x_2^{(i)}, \ldots, x_{n_i}^{(i)}\}$ is an independent set of $n_i \geq 1$ eigenvectors of A corresponding to the eigenvalue λ_i. Show that the union of the sets $\{x_1^{(1)}, x_2^{(1)}, \ldots, x_{n_1}^{(1)}\} \cup \cdots \cup \{x_1^{(k)}, x_2^{(k)}, \ldots, x_{n_k}^{(k)}\}$ is an independent set. *Hint:* If some linear combination is zero, say

$$0=\sum_{i=1}^{k}\sum_{j=1}^{n_i} c_{ij}x_j^{(i)}=\sum_{i=1}^{k} y^{(i)}$$

use (1.3.8) to show that each $y^{(i)}=0$.

11. Provide the details for the following alternate, and more constructive, proof of Lemma (1.3.17): (a) Show that if $A, B\in M_n$ commute, then they have a common eigenvector. *Hint:* Let x be an eigenvector of A, $Ax=\lambda x$, $x\neq 0$, and consider the sequence $x, Bx, B^2x, B^3x, \ldots$. There must be a first element of this sequence that is dependent upon its predecessors, say B^kx, so $S=\text{Span}\{x, Bx, B^2x, \ldots, B^{k-1}x\}$ is a subspace invariant under B and hence there is some nonzero $y\in S$ with $By=\mu y$. But $AB^jx=B^jAx=B^j\lambda x=\lambda B^jx$, so every vector in S is an eigenvector for A as well. (b) If $\mathfrak{F}=\{A_1, A_2, \ldots, A_m\}$ is a finite commuting family, use induction to show that there is a common eigenvector for all A_i. *Hint:* If $y\neq 0$ is a common eigenvector for $A_1, A_2, \ldots, A_{m-1}$, consider the sequence $y, A_my, A_m^2y, A_m^3y, \ldots$ as in (a). (c) If $\mathfrak{F}\subset M_n$ is a commuting family that does not have finite cardinality, observe that there cannot be more than n^2 linearly independent matrices in $\mathfrak{F}$. Select a maximal independent set and use (b); argue that a common eigenvector for this finite set is a common eigenvector for all of $\mathfrak{F}$.

12. If $\Lambda=\text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)\in M_n$ has n distinct diagonal entries, use the ideas from the proof of Theorem (1.3.12) to show that $\Lambda B=B\Lambda$ for some $B\in M_n$ if and only if B is itself diagonal (but not necessarily with distinct diagonal entries).

13. Suppose $A\in M_n$ has n distinct eigenvalues. If $AB=BA$ for some $B\in M_n$, show that B is diagonalizable and that A and B are simultaneously diagonalizable. *Hint:* If $A=S\Lambda S^{-1}$ with Λ diagonal, show that Λ commutes with $S^{-1}BS$ and use Problem 12.

14. Extend the result of Problem 13 to a commuting family $\mathfrak{F}\subset M_n$ that contains at least one matrix with n distinct eigenvalues. Compare this result to Theorem (1.3.19), which assumes that all the members of the family are diagonalizable. Is this a stronger result?

15. Consider the block diagonal matrix $\Lambda=\text{diag}(\lambda_1I_1, \lambda_2I_2, \ldots, \lambda_kI_k)\in M_n$ with $I_j\in M_{n_j}$, $\lambda_i\neq\lambda_j$ if $i\neq j$, and $n_1+n_2+\cdots+n_k=n$. Show that $\Lambda B=B\Lambda$ for some $B\in M_n$ if and only if the matrix B has the block diagonal form $B=\text{diag}(B_1, B_2, \ldots, B_k)$ with $B_j\in M_{n_j}$, $j=1, 2, \ldots, k$. How is this result related to Problem 12?

16. Let $A, B\in M_n$, and suppose that either A or B is nonsingular. If AB is diagonalizable, show that BA is also diagonalizable. Consider $A=\begin{bmatrix}0&1\\0&0\end{bmatrix}$ and $B=\begin{bmatrix}1&1\\0&0\end{bmatrix}$ to show that this need not be true if both A and B are singular.

1.4 Eigenvectors

Thus far, the eigenvalues of $A\in M_n$ have been emphasized relative to the eigenvectors. The eigenvectors are also important not only for their role in diagonalizability, but also for their utility in a variety of applications. We discuss them somewhat further here, but we begin with an additional observation about eigenvalues.

1.4.1 **Observation.** Let $A\in M_n$. (a) The eigenvalues of A^T are the same as those of A, counting multiplicities. (b) The eigenvalues of A^* are the complex conjugates of the eigenvalues of A, counting multiplicities.

Proof: Since $\det(tI-A^T)=\det(tI-A)^T=\det(tI-A)$, $p_{A^T}(t)=p_A(t)$, and (a) follows. Similarly, $\det(\bar{t}I-A^*)=\det[(tI-A)^*]=\overline{\det(tI-A)}$, which implies $p_{A^*}(\bar{t})=\overline{p_A(t)}$, and (b) follows. □

Exercise. If $x, y\in \mathbf{C}^n$ are both eigenvectors of $A\in M_n$ corresponding to the eigenvalue λ, show that any nonzero linear combination of x and y is also an eigenvector corresponding to λ. Conclude that the set of all eigenvectors associated with a particular $\lambda\in\sigma(A)$, together with the 0 vector, is a subspace of $\mathbf{C}^n$.

Exercise. Observe that the subspace described in the preceding exercise is precisely the null space of $A-\lambda I$.

1.4.2 **Definition.** Let $A\in M_n$. For a given $\lambda\in\sigma(A)$, the set of all vectors $x\in\mathbf{C}^n$ satisfying $Ax=\lambda x$ is called the *eigenspace* of A corresponding to the eigenvalue λ. Note that every nonzero element of this eigenspace is an eigenvector of A corresponding to λ.

Exercise. Show that an eigenspace of A corresponding to an eigenvalue λ is an A-invariant subspace, but not conversely. Show that a *minimal* A-invariant subspace (containing no strictly lower-dimensional, nontrivial A-invariant subspace) is the span of a single eigenvector of A. *Hint:* Use the exercise preceding (1.3.17).

If one knows an eigenvalue of $A\in M_n$, a conceptually simple, but not necessarily practical, approach to computing an associated eigenvector is to solve the linear system

$$(A-\lambda I)x=0$$

The set of all solutions constitutes the eigenspace.

1.4.3 **Definition.** The dimension of the eigenspace of $A \in M_n$ corresponding to the eigenvalue λ is called the *geometric multiplicity* of the eigenvalue λ. The multiplicity of λ as a zero of the characteristic polynomial $p_A(\cdot)$ (the notion of multiplicity to which we have referred thus far) is called the *algebraic multiplicity* of the eigenvalue λ. In general, these two concepts are different. If the term *multiplicity* is used without qualification in reference to an eigenvalue, it usually means the algebraic multiplicity. We shall follow this convention.

Notice that the geometric multiplicity is just the maximum number of linearly independent eigenvectors associated with an eigenvalue.

Exercise. Show that the geometric multiplicity of an eigenvalue λ of $A \in M_n$ is never more, and can be less, than its algebraic multiplicity. If the algebraic multiplicity is at least 1, then the geometric multiplicity is at least 1. *Hint:* Suppose the geometric multiplicity of λ is k, and let $S \in M_n$ be nonsingular with its first k columns linearly independent eigenvectors of A corresponding to λ. Use reasoning similar to that used in (1.3.7) to show that $S^{-1}AS$ has the form $\left[\begin{array}{c|c} \lambda I & * \\ \hline 0 & * \end{array}\right]$, $I \in M_k$, and conclude that the algebraic multiplicity of λ is at least k.

1.4.4 **Definitions.** A matrix $A \in M_n$, some eigenvalue of which has strictly smaller geometric than algebraic multiplicity, is said to be *defective*. If the geometric multiplicity is the same as the algebraic multiplicity for each eigenvalue, A is said to be *nondefective*. If each eigenvalue of $A \in M_n$ has geometric multiplicity exactly 1 (regardless of the algebraic multiplicity), A is called *nonderogatory*. All of these definitions are classical and enjoy somewhat limited current usage.

Notice that a nonderogatory, nondefective matrix is simply a matrix with distinct eigenvalues. Also, a matrix $A \in M_n$ is diagonalizable if and only if A is nondefective. This is just a restatement of (1.3.7) that emphasizes the necessity of the existence of enough linearly independent eigenvectors associated with each eigenvalue.

1.4.5 **Example.** Even though A and A^T have the same eigenvalues, their eigenvectors corresponding to a given eigenvalue may be very different. For example, let

$$A=\begin{bmatrix} 2 & 3 \\ 0 & 4 \end{bmatrix}$$

Then the (one-dimensional) eigenspace of A corresponding to the eigenvalue 2 is spanned by $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$, while the corresponding eigenspace of A^T is spanned by $\begin{bmatrix} 1 \\ -3/2 \end{bmatrix}$.

Exercise. Verify the details of (1.4.5).

It should be clear that the theory of eigenvalues and eigenvectors we have developed thus far could have been developed in parallel for left multiplication by row vectors. The eigenvalues would be the same, but the eigenvectors would, in general, be different (even allowing for rows vs. columns).

1.4.6 **Definition.** A nonzero vector $y \in \mathbf{C}^n$ is called a *left eigenvector* of $A \in M_n$ corresponding to $\lambda \in \sigma(A)$ if

$$y^*A=\lambda y^*$$

If necessary for clarity, we refer to the vector x of (1.1.3) as a *right eigenvector.* When the context does not require distinction, we just say *eigenvector.*

Exercise. Show that a left eigenvector y corresponding to the eigenvalue λ of $A \in M_n$ is a right eigenvector of A^* corresponding to $\bar{\lambda}$, and also that $\bar{y}$ is a right eigenvector of A^T corresponding to λ. Show by example that, even for $A \in M_n(\mathbf{R})$, right and left eigenvectors need not be the same.

Recall from (0.6.2) that two vectors $x, y \in \mathbf{C}^n$ are called *orthogonal* if $y^*x=0$. The following result is known as the *principle of biorthogonality.*

1.4.7 **Theorem.** If $A \in M_n$ and if $\lambda, \mu \in \sigma(A)$, with $\lambda \neq \mu$, then any left eigenvector of A corresponding to μ is orthogonal to any right eigenvector of A corresponding to λ.

Proof: Let $y \in \mathbf{C}^n$ be a left eigenvector of A corresponding to μ and let $x \in \mathbf{C}^n$ be a right eigenvector of A corresponding to λ. Manipulate y^*Ax in two ways:

$$\begin{aligned} y^*Ax &= y^*(\lambda x)=\lambda(y^*x) \\ &= (\mu y^*)x=\mu(y^*x) \end{aligned}$$

Since $\lambda \neq \mu$, the only possible way to have $\lambda y^*x = \mu y^*x$ is to have $y^*x = 0$; that is, x and y are orthogonal. □

Exercise. If $A^* = A \in M_n$, that is, A is *Hermitian*, and if A has distinct eigenvalues, show that there are n pair-wise orthogonal (right) eigenvectors of A. Recall from Problem 8 of Section (1.1) that the eigenvalues of A are all real. *Hint:* Since $A^* = A$, left eigenvectors coincide with right eigenvectors. Apply (1.4.7).

We shall see in the next chapter that the assumption of distinct eigenvalues is unnecessary in the statement of the above exercise.

We next note that eigenvectors transform under similarity in a simple way. The eigenvalues are, of course, unchanged by similarity.

1.4.8 **Theorem.** Let $A, B \in M_n$. If $x \in \mathbf{C}^n$ is an eigenvector corresponding to $\lambda \in \sigma(B)$ and if B is similar to A via S, then Sx is an eigenvector of A corresponding to the eigenvalue λ.

Proof: If $B = S^{-1}AS$ and if $Bx = \lambda x$, then $S^{-1}ASx = \lambda x$, or $ASx = \lambda Sx$. Since S is nonsingular and $x \neq 0$, $Sx \neq 0$, and hence Sx is an eigenvector of A. □

Exercise. Verify that $e = [1, 1, 1]^T$ is an eigenvector of

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \\ 2 & 3 & 1 \end{bmatrix}$$

If $D = \text{diag}(1, 2, 3)$, determine an entry-wise positive eigenvector of $D^{-1}AD$.

As a final observation in this section, we note that eigenvectors can be used to gain information about eigenvalues of principal submatrices. This information yields another proof of the inequality between the geometric and algebraic multiplicities of an eigenvalue.

1.4.9 **Theorem.** Let $A \in M_n$ and $\lambda \in \mathbf{C}$ be given, and let $k \geq 1$ be a given positive integer. Consider the following three statements:

(a) λ is an eigenvalue of A of geometric multiplicity at least k.
(b) If $\hat{A} \in M_m$ is a principal submatrix of A and if $m > n-k$, then λ is an eigenvalue of $\hat{A}$.
(c) λ is an eigenvalue of A of algebraic multiplicity at least k.

Then (a) implies (b), and (b) implies (c). In particular, the algebraic multiplicity of an eigenvalue is at least as great as its geometric multiplicity.

Proof: Assume (a) and let $\hat{A} \in M_m$ be a principal submatrix of A with $m > n-k$. Using a permutation similarity and (1.4.8), there is no loss of generality to assume that $\hat{A}$ appears in the upper left corner of A. Let $v_1, \ldots, v_k$ be linearly independent eigenvectors of A corresponding to the eigenvalue λ. Partition A and each vector v_i as

$$A = \begin{bmatrix} \hat{A} & * \\ * & * \end{bmatrix}, \qquad \hat{A} \in M_m;$$

$$v_i = \begin{bmatrix} u_i \\ w_i \end{bmatrix}, \qquad u_i \in \mathbf{C}^m, \quad w_i \in \mathbf{C}^{n-m}, \quad i = 1, 2, \ldots, k$$

The vectors $w_1, \ldots, w_k$ are dependent because they are k vectors in a space of dimension $n-m < n-(n-k) = k$, so there are scalars $\alpha_1, \ldots, \alpha_k \in \mathbf{C}$, not all zero, such that $\alpha_1 w_1 + \cdots + \alpha_k w_k = 0$. Then $v \equiv \alpha_1 v_1 + \cdots + \alpha_k v_k = \left[\begin{smallmatrix} u \\ 0 \end{smallmatrix}\right] \neq 0$, where $u = \alpha_1 u_1 + \cdots + \alpha_k u_k \neq 0$ and $Av = \lambda v$. Writing this equation in partitioned form gives

$$Av = \begin{bmatrix} \hat{A} & * \\ * & * \end{bmatrix} \begin{bmatrix} u \\ 0 \end{bmatrix} = \begin{bmatrix} \hat{A}u \\ * \end{bmatrix} = \lambda v = \begin{bmatrix} \lambda u \\ 0 \end{bmatrix}$$

This shows that λ is an eigenvalue of $\hat{A}$, which is the assertion in (b).

Now assume (b) and recall the identity (1.2.13), which relates the derivative of the characteristic polynomial $p_A(t)$ to the characteristic polynomials $p_{A_i}(t)$ of the n principal submatrices $A_1, \ldots, A_n$ of A. If $k = 1$, there is nothing to prove. If $k > 1$, then (b) says that λ is an eigenvalue of each A_i and hence $p_{A_i}(\lambda) = 0$ and $p'_A(\lambda) = 0$. If $k > 2$, differentiate the identity (1.2.13) to get

$$p''_A(t) = \sum_{i=1}^{n} p'_{A_i}(t) \tag{1.4.10}$$

and use (1.2.13) to replace each derivative on the right-hand side with a sum of characteristic polynomials of principal submatrices of each A_i. Since a principal submatrix of A_i, with one row and column deleted, is a principal submatrix of A of size $n-2$, the assumption in (b) and the identity (1.2.13) applied to each A_i permit us to conclude that $p''_A(\lambda) = 0$. Repetition of this argument shows that the successive derivatives $p_A^{(i)}(\lambda)$ vanish for $i = 0, 1, \ldots, k-1$, and hence λ has algebraic multiplicity at least k. □

Problems

1. Show that $A \in M_n$ has rank 1 if and only if there exist two nonzero vectors $x, y \in \mathbf{C}^n$ such that

$$A = xy^*$$

Show that (a) such an A has at most one nonzero eigenvalue (of algebraic multiplicity 1); (b) this eigenvalue is y^*x; and (c) x is a right and y is a left eigenvector corresponding to this eigenvalue. What is the geometric multiplicity of the eigenvalue 0?

2. Show that a matrix $A \in M_n$ of rank k may be written as

$$A = x^{(1)}y^{(1)*} + \cdots + x^{(k)}y^{(k)*},$$

for $x^{(i)}, y^{(i)} \in \mathbf{C}^n$, $i = 1, \ldots, k$, that is, as a sum of k rank 1 matrices. *Hint:* Find k linearly independent rows and columns and use the fact that the others can be written in terms of these.

3. Suppose that $T \in M_n$ is upper triangular with distinct eigenvalues $t_{11}, \ldots, t_{nn}$ occurring from upper left to lower right, down the diagonal. Show that there is a right eigenvector of T corresponding to t_{ii} whose last $n-i$ components are all 0, and a left eigenvector of T corresponding to t_{ii} whose first $i-1$ components are all 0. What if the t_{ii} are not distinct?

4. Show that the (only) eigenvalue 1 of the matrix displayed in (1.2.7b) has geometric multiplicity 1. Describe the associated eigenspace.

5. Consider the block triangular matrix

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \qquad A_{ii} \in M_{n_i}, \quad i = 1, 2$$

Show that the eigenvalues of A are those of A_{11} together with those of A_{22}, counting multiplicities. If $x \in \mathbf{C}^{n_1}$ is a right eigenvector of A_{11} corresponding to $\lambda \in \sigma(A_{11})$, and if $y \in \mathbf{C}^{n_2}$ is a left eigenvector of A_{22} corresponding to $\mu \in \sigma(A_{22})$, show that $\begin{bmatrix} x \\ 0 \end{bmatrix} \in \mathbf{C}^{n_1+n_2}$ is a right eigenvector and $\begin{bmatrix} 0 \\ y \end{bmatrix}$ is a left eigenvector of A corresponding to λ and μ, respectively. What can you say about left and right eigenvectors of A corresponding to λ and μ, respectively? Can you generalize these observations to block triangular matrices with arbitrarily many diagonal blocks?

6. If $A \in M_n$ has component-wise positive left and right eigenvectors corresponding to an eigenvalue of geometric multiplicity 1, show that A has no other component-wise nonnegative eigenvectors, except for multiples of these.

7. In this problem we outline the *power method* for finding the largest eigenvalue and an associated eigenvector of $A \in M_n$. We make some simplifying assumptions and allude to analytical details that can be made precise. Suppose that $A \in M_n$ has distinct eigenvalues $\lambda_1, \ldots, \lambda_n$ and that

there is exactly one eigenvalue λ_n of maximum modulus $\rho(A)$. If $x^{(0)} \in \mathbf{C}^n$ is *not* orthogonal to a left eigenvector associated with λ_n, show that the sequence

$$x^{(k+1)} = \frac{1}{(x^{(k)*}x^{(k)})^{1/2}} Ax^{(k)}, \quad k = 0, 1, 2, \ldots$$

approaches an eigenvector of A and the ratios of a given nonzero component in the vectors $Ax^{(k)}$ and $x^{(k)}$ approach λ_n. *Hint:* Assume without loss of generality that $\lambda_n = 1$ and let $y^{(1)}, \ldots, y^{(n)}$ be linearly independent eigenvectors corresponding to $\lambda_1, \ldots, \lambda_n$. The vector $x^{(0)}$ may be written uniquely as

$$x^{(0)} = \alpha_1 y^{(1)} + \cdots + \alpha_n y^{(n)}$$

with $\alpha_n \neq 0$. Notice that $x^{(k)} = \alpha_1 \lambda_1^k y^{(1)} + \cdots + \alpha_n \lambda_n^k y^{(n)}$, except for a factor of scale. Since $|\lambda_i| < 1$, $|\lambda_i|^k \to 0$, $i = 1, \ldots, n-1$ and this sum approaches a multiple of $y^{(n)}$.

8. Further eigenvalues (and eigenvectors) can be calculated using the power method via a bridge, called *deflation,* which delivers a square matrix, of size 1 smaller, whose eigenvalues are the remaining eigenvalues of $A \in M_n$. Let λ_n and $y^{(n)}$ be an eigenvalue and eigenvector of A (calculated using the power method or otherwise), and let $S \in M_n$ be nonsingular with first column $y^{(n)}$. Show that

$$S^{-1}AS = \left[\begin{array}{c|c} \lambda_n & * \\ \hline 0 & A_1 \end{array}\right]$$

and that the eigenvalues of $A_1 \in M_{n+1}$ are $\lambda_1, \ldots, \lambda_{n-1}$ in the notation of Problem 7. Another eigenvalue may be calculated from A_1 and the deflation repeated – and so on.

9. Let $A \in M_n$ have eigenvalues $\lambda_1, \ldots, \lambda_{n-1}, 0$, so that rank $A \leq n-1$, and suppose that the last row of A is a linear combination of the others. (a) If A is partitioned as

$$\begin{bmatrix} A_{11} & a_{12} \\ a_{21}^T & a_{22} \end{bmatrix}$$

in which $A_{11} \in M_{n-1}$, show that there is a vector $b \in \mathbf{C}^{n-1}$ such that

$$a_{21}^T = b^T A_{11} \quad \text{and} \quad a_{22} = b^T a_{12}$$

Interpret b in terms of a left eigenvector of A corresponding to 0. (b) Show also that $A_{11} + a_{12}b^T \in M_{n-1}$ has eigenvalues $\lambda_1, \ldots, \lambda_{n-1}$. *Hint:* Consider similarity of A via

$$\begin{bmatrix} I & 0 \\ b^T & 1 \end{bmatrix}$$

Notice that this is another version of deflation since a matrix of smaller size with the remaining eigenvalues is produced. If one knows one eigenvalue λ of A, then the process described in this problem can be applied to $P(A-\lambda I)P^{-1}$, for a suitable permutation P.

10. Let $T \in M_n$ be a nonsingular matrix whose columns are left eigenvectors of $A \in M_n$. Show that the columns of $(T^*)^{-1}$ are right eigenvectors of A.

CHAPTER 2

Unitary equivalence and normal matrices

We next study a special type of similarity that is intimately involved with many aspects of the application of matrix analysis.

2.0 Introduction

For a general nonsingular matrix $S \in M_n$, we made an initial study of similarity via S in Chapter 1. For certain very special nonsingular matrices, called unitary matrices, the inverse of S has a simple form: $S^{-1} = S^*$. Similarity of $A \in M_n$ via a unitary matrix, $A \to S^*AS$, is not only conceptually simpler (S^* is much easier to evaluate than S^{-1}) than general similarity, but it has a number of attractive features that will become clearer through the development to follow. As a general rule, unitary similarities are preferable to general similarities, and it is therefore useful to know what can be achieved through unitary similarity. Equivalence classes under unitary similarity are, however, finer than under general similarity (two matrices can be similar but not unitarily similar), and correspondingly less can be achieved. For this reason, we shall return to study general similarity further in Chapter 3.

The transformation $A \to S^*AS$, $A \in M_n$, in which S is assumed to be nonsingular but not necessarily unitary, is called **congruence* and will be studied in Chapter 4. This transformation, too, is an equivalence relation on M_n with a number of attractive features (different from those of similarity). It is important to realize that similarity by a unitary matrix is *both* a similarity and a *congruence and this is the broadest class of transformations that shares the properties of both.

2.1 Unitary matrices

2.1.1 **Definition.** Recall that the vectors $x_1, \ldots, x_k \in \mathbf{C}^n$ form an *orthogonal set* if $x_i^* x_j = 0$ for all pairs $1 \le i < j \le k$. If, in addition, the vectors are normalized, $x_i^* x_i = 1$, $i = 1, \ldots, k$, then the set is called *orthonormal.*

Exercise. If $\{y_1, \ldots, y_k\}$ is an orthogonal set of nonzero vectors, show that the set $\{x_1, \ldots, x_k\}$ defined by $x_i = (y_i^* y_i)^{-1/2} y_i$, $i = 1, \ldots, k$, is an orthonormal set.

2.1.2 **Theorem.** An orthonormal set of vectors is linearly independent.

Proof: Suppose that $\{x_1, \ldots, x_k\}$ is an orthonormal set, and suppose $0 = \alpha_1 x_1 + \cdots + \alpha_k x_k$. Then $0 = 0^* 0 = \sum_{i,j} \bar{\alpha}_i \alpha_j x_i^* x_j = \sum_{i=1}^k |\alpha_i|^2 x_i^* x_i$ because the vectors x_i are orthogonal, and $\sum_{i=1}^k |\alpha_i|^2 x_i^* x_i = \sum_{i=1}^k |\alpha_i|^2 = 0$ because the vectors x_i are normalized. Thus, all $\alpha_i = 0$ and hence $\{x_1, \ldots, x_k\}$ is a linearly independent set. □

Exercise. Show that an orthogonal set of nonzero vectors is linearly independent.

Exercise. Show that if $x_1, \ldots, x_k \in \mathbf{C}^n$ is an orthogonal set, then either $k \le n$ *or* at least $k - n$ of the vectors x_i are equal to zero.

An independent set need not be orthonormal, of course, but one can apply the Gram–Schmidt orthonormalization procedure (0.6.4) to it and obtain an orthonormal set with the same span as the original set.

Exercise. Show that any k-dimensional real or complex vector space has an orthonormal basis (a basis consisting of an orthonormal set).

2.1.3 **Definition.** A matrix $U \in M_n$ is said to be *unitary* if $U^* U = I$. If, in addition, $U \in M_n(\mathbf{R})$, U is said to be *real orthogonal.*

The unitary matrices in M_n form a remarkable and important set. We list some of the basic equivalent conditions for U to be unitary in (2.1.4).

Exercise. If $A \in M_n$ is given and $BA = I$ for some $B \in M_n$, show that A is nonsingular, B is unique, and $AB = I$. We write $B = A^{-1}$. *Hint:* If $Ax = 0$, then $x = Ix = BAx$. Nonsingularity implies that the equations $Ax = y$ and $x^T A = y^T$ each have a unique solution for any given $y \in \mathbf{C}^n$. Argue (column by column) that $AB_R = I$ and (row by row) that $B_L A = I$ have unique solutions $B_L, B_R \in M_n$. Now calculate $B_L A B_R$ in two ways to show that $B_L = B_R$.

2.1.4 **Theorem.** If $U \in M_n$, the following are equivalent:

(a) U is unitary;
(b) U is nonsingular and $U^* = U^{-1}$;
(c) $UU^* = I$;
(d) U^* is unitary;
(e) The columns of U form an orthonormal set;
(f) The rows of U form an orthonormal set; and
(g) For all $x \in \mathbf{C}^n$, the Euclidean length of $y \equiv Ux$ is the same as that of x; that is, $y^* y = x^* x$.

Proof: Statement (a) implies (b) since U^{-1} (when it exists) is that unique matrix, left multiplication by which produces I; the definition of unitary guarantees that U^* is such a matrix. Since $BA = I$ if and only if $AB = I$ (for $A, B \in M_n$), (b) implies (c). Since $(U^*)^* = U$, (c) implies that U^* satisfies the requirement necessary to be unitary; that is, (c) implies (d). Since the converse of each of these implications is similarly observed, (a)–(d) are equivalent.

Considering the mechanics of matrix multiplication and letting $u^{(i)}$ denote the ith column of U, $i = 1, \ldots, n$, the statement $U^* U = I$ means that

$$u^{(i)*} u^{(j)} = \begin{cases} 0 & \text{if } j \ne i \\ 1 & \text{if } j = i \end{cases}$$

Thus, $U^* U = I$ is another way of saying that the columns of U are orthonormal, and (a) is equivalent to (e). Similarly, (d) and (f) are equivalent.

If (a) holds and $y = Ux$, then $y^* y = x^* U^* U x = x^* I x = x^* x$, so that (a) implies (g). To verify the converse, on the other hand, requires somewhat more elaborate calculation. Tools occurring later in this book would make this fact more immediate, however. First consider the case $n = 2$. Assuming (g) and letting $x = \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$, we find that $1 = x^* x = y^* y = x^* U^* U x =$ the 1, 1 entry of $U^* U$. Similarly, letting $x = \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]$, we conclude that the 2, 2 entry of $U^* U$ is also 1, and $U^* U$ must have the form

$$\begin{bmatrix} 1 & a \\ \bar{a} & 1 \end{bmatrix}$$

where a is the inner product of column 1 and column 2 of U, and $\bar{a}$ is the inner product of column 2 and column 1. Letting $x = \left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$ in (g) and again calculating, we find $2 = x^* x = y^* y = x^* U^* U x = 2 + (a + \bar{a})$. Letting $x = \left[\begin{smallmatrix} 1 \\ i \end{smallmatrix}\right]$,

we find $2=2+i(a-\bar{a})$. Thus, $a+\bar{a}=2\operatorname{Re} a=0$ and $a-\bar{a}=2i\operatorname{Im} a=0$, and hence $a=0$. This means that if $x^*U^*Ux=x^*x$ for all $x\in\mathbf{C}^2$, then $U^*U=I$; that is, U is unitary (if $U\in M_2$). Now consider $n>2$, and let $A=U^*U$. Let $x\in\mathbf{C}^n$ be such that all components other than the ith and jth, $i\le j$, are 0. Then

$$x^*Ax=[\bar{x}_i,\bar{x}_j]A(\{i,j\})\begin{bmatrix} x_i \\ x_j \end{bmatrix}$$

[see (0.7.1) for submatrix notation], and we have just shown that (g) implies that $A(\{i,j\})=I\in M_2$. Since i and j are arbitrary, we conclude that every 2-by-2 principal submatrix of A is the 2-by-2 identity. The only such A is $A=I\in M_n$, and, since the case $n=1$ is obvious, we conclude that (g) implies (a), which completes the proof. □

2.1.5 **Definition.** A linear transformation $T\colon \mathbf{C}^n\to\mathbf{C}^m$ is called a *Euclidean isometry* if $x^*x=(Tx)^*(Tx)$ for all $x\in\mathbf{C}^n$. Theorem (2.1.4) says that a square complex matrix $U\in M_n$ is a Euclidean isometry (via $U\colon x\to Ux$) if and only if it is unitary. See Section 5.2 for other kinds of isometries.

Exercise. Let

$$T(\theta)=\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

where θ is a real parameter. (a) If $U\in M_2(\mathbf{R})$, show that U is real orthogonal if and only if $U=T(\theta)$ or

$$U=\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}T(\theta)$$

for some $\theta\in\mathbf{R}$. (b) If $U\in M_2(\mathbf{R})$, show that U is real orthogonal if and only if $U=T(\theta)$ or

$$U=\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}T(\theta)$$

for some $\theta\in\mathbf{R}$. These are two different presentations, in terms of a parameter θ, of the 2-by-2 real orthogonal matrices. Interpret them geometrically.

2.1.6 **Observation.** If $U,V\in M_n$ are unitary (respectively real orthogonal), then the product UV is also unitary (respectively real orthogonal).

Exercise. Use (b) of (2.1.4) to prove (2.1.6).

Exercise. If $\{x_1,x_2,\ldots,x_k\}\subseteq\mathbf{C}^n$ is an orthonormal set and if $U\in M_n$ is unitary, show that $\{Ux_1,\ldots,Ux_k\}$ is an orthonormal set.

2.1.7 **Observation.** The set of unitary (respectively real orthogonal) matrices in M_n forms a group. This group is generally referred to as the *n-by-n unitary* (respectively *orthogonal*) *group*, a subgroup of $GL(n,\mathbf{C})$ [see Section (0.5)].

Exercise. Recall that a *group* is a set that is *closed* under a single *associative* binary operation ("multiplication") and such that the *identity* for and *inverses* under the operation are contained in the set. Verify (2.1.7). *Hint:* Use (2.1.6) for closure; matrix multiplication is associative; $I\in M_n$ is unitary; and $U^*=U^{-1}$ is again unitary.

The set (group) of unitary matrices in M_n has another very important property. Notions of "convergence" and "limit" of a sequence of matrices will be presented precisely in Chapter 5, but can be understood here in terms of "convergence" and "limit" of each i,j entry. The defining identity $U^*U=I$ means that every column of U has Euclidean length 1, and hence no entry u_{ij} of $U=[u_{ij}]$ can have absolute value greater than 1. If we think of the set of unitary matrices as a subset of $\mathbf{C}^{n^2}$, this says it is a *bounded* subset. If $U_k\equiv[u_{ij}^{(k)}]$ is a sequence of unitary matrices, $k=1,2,\ldots$ such that $\lim_{k\to\infty}u_{ij}^{(k)}\equiv u_{ij}^{(0)}$ exists for all $i,j=1,2,\ldots,n$, then from the identity $U_k^*U_k=I$ for all $k=1,2,\ldots$ we see that $\lim_{k\to\infty}U_k^*U_k=U_0^*U_0=I$, where $U_0\equiv[u_{ij}^{(0)}]$. Thus, the limit matrix U_0 is also unitary. This says that the set of unitary matrices is a *closed* subset of $\mathbf{C}^{n^2}$.

Since a closed and bounded subset of a finite dimensional Euclidean space is a *compact* set (see Appendix E), we conclude that the set (group) of unitary matrices in M_n is compact. For our purposes, the most important consequence of this observation is the following *selection principle* for unitary matrices.

2.1.8 **Lemma.** Let $U_1,U_2,\ldots$ be a given sequence of unitary matrices in M_n. There exists a subsequence $U_{k_1},U_{k_2},\ldots$ such that all of the entries of U_{k_i} converge (as sequences of complex numbers) to the entries of a unitary matrix U_0 as $i\to\infty$.

Proof: All that is required here is the fact that from any infinite sequence in a compact set one may always select a convergent subseqence. We have already observed that if a sequence of unitary matrices converges to some matrix, then the limit matrix must be unitary. □

The unitary limit guaranteed by the lemma need not be unique; it can depend upon the subsequence chosen.

Exercise. Consider the sequence of unitary matrices

$$U_k = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}^k, \qquad k = 1, 2, \ldots$$

Show that there are two possible limits of subsequences.

Exercise. The selection principle (2.1.8) applies as well to the orthogonal group; that is, a sequence of real orthogonal matrices has a subsequence that converges to a real orthogonal matrix. Verify this by tracing through the same logic in the real case.

Compactness of the unitary group is invoked in Problem 3 of the next section. We shall have occasion to use it elsewhere in the book.

A unitary matrix U has the property that U^{-1} *equals* U^*. One way to generalize the notion of a unitary matrix is to require that U^{-1} be *similar* to U^*. The set of such matrices is easily characterized as the range of the mapping $A \to A^{-1}A^*$ for all nonsingular $A \in M_n$.

2.1.9 **Theorem.** Let $A \in M_n$ be a nonsingular matrix. Then A^{-1} is similar to A^* if and only if there is a nonsingular matrix $B \in M_n$ such that $A = B^{-1}B^*$.

Proof: If $A = B^{-1}B^*$ for some nonsingular $B \in M_n$, then $A^{-1} = (B^*)^{-1}B$ and $B^*A^{-1}(B^*)^{-1} = B(B^*)^{-1} = (B^{-1}B^*)^* = A^*$, so A^{-1} is similar to A^* via the similarity B^*. Conversely, if A^{-1} is similar to A^*, then there is a nonsingular matrix $S \in M_n$ such that $SA^{-1}S^{-1} = A^*$. Set $S_\theta \equiv e^{i\theta}S$ for $\theta \in \mathbf{R}$ and notice that $S_\theta A^{-1}S_\theta^{-1} = e^{i\theta}SA^{-1}(e^{-i\theta}S^{-1}) = SA^{-1}S^{-1} = A^*$. But then $S_\theta = A^*S_\theta A$ and $S_\theta^* = A^*S_\theta^*A$. Adding these two identities gives $H_\theta = A^*H_\theta A$, where $H_\theta \equiv S_\theta + S_\theta^*$ is Hermitian. If H_θ is singular, there is some nonzero $x \in \mathbf{C}^n$ such that $0 = H_\theta x = S_\theta x + S_\theta^* x$, so $-x = S_\theta^{-1}S_\theta^* x = e^{-2i\theta}S^{-1}S^*x$ and $S^{-1}S^*x = -e^{2i\theta}x$. Choose a value of $\theta = \theta_0 \in [0, 2\pi)$ such that $-e^{2i\theta_0}$ is not an eigenvalue of $S^{-1}S^*$; the resulting Hermitian matrix $H \equiv H_{\theta_0}$ is nonsingular and has the property that $H = A^*HA$.

Now choose any complex α such that $|\alpha| = 1$ and α is not an eigenvalue of A^*. Set $B \equiv \beta(\alpha I - A^*)H$, where the complex parameter $\beta \neq 0$ is to be chosen, and observe that B is nonsingular. We want to have $A = B^{-1}B^*$, or $BA = B^*$. Compute $B^* = H(\bar{\beta}\bar{\alpha}I - \bar{\beta}A)$, and $BA = \beta(\alpha I - A^*)HA = \beta(\alpha HA - A^*HA) = \beta(\alpha HA - H) = H(\alpha\beta A - \beta I)$. We shall be done if we can select a nonzero β such that $\beta = -\bar{\beta}\bar{\alpha}$, but if $\alpha = e^{i\psi}$, then $\beta = e^{i(\pi - \psi)/2}$ will do. □

Problems

1. If $U \in M_n$ is unitary, show that $|\det U| = 1$.

2. If $\lambda \in \sigma(U)$ and $U \in M_n$ is unitary, show that $|\lambda| = 1$. *Hint:* Use the isometry property (2.1.4g).

3. Given real parameters $\theta_1, \theta_2, \ldots, \theta_n$, show that

$$U = \operatorname{diag}(e^{i\theta_1}, e^{i\theta_2}, \ldots, e^{i\theta_n})$$

is unitary.

4. Characterize the diagonal real orthogonal matrices.

5. Show that the permutation matrices (0.9.5) in M_n are orthogonal and that the permutation matrices form a subgroup (a subset which is itself a group) of the group of real orthogonal matrices. How many different permutation matrices are there in M_n?

6. Can you give a presentation in terms of parameters of the 3-by-3 orthogonal group? Recall the two presentations of the 2-by-2 orthogonal group given in the section.

7. Provide the details for the following alternate proof that (g) implies (a) in (2.1.4). Show that (g) means that $x^*(U^*U - I)x = 0$ for all $x \in \mathbf{C}^n$. Let $H \equiv U^*U - I$ and observe that $H = H^*$. Consider $0 = (x + e^{i\theta}y)^*H(x + e^{i\theta}y)$ for all $x, y \in \mathbf{C}^n$ and all $\theta \in \mathbf{R}$. Expand this identity and show that $x^*Hy = 0$ for all $x, y \in \mathbf{C}^n$. Conclude that $H = 0$ by making systematic choices for x and y.

8. A matrix $A \in M_n$ such that $AA^T = I$ is said to be *orthogonal.* A real orthogonal matrix is unitary, but a nonreal orthogonal matrix need not be unitary. (a) Let

$$K = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \in M_2(\mathbf{R})$$

Show that $A(t) = (\cosh t)I + (i \sinh t)K \in M_2$ is orthogonal for all $t \in \mathbf{R}$ but that $A(t)$ is unitary only for $t = 0$. The hyperbolic functions are defined by $\cosh t = (e^t + e^{-t})/2$, $\sinh t = (e^t - e^{-t})/2$. (b) Show that, unlike the unitary matrices, the set of complex orthogonal matrices is not a bounded set, and it is therefore not a compact set. (c) Show that, like the unitary matrices, the set of complex orthogonal matrices of a given size forms a group. Despite this fact, common usage reserves the term *orthogonal group* for the smaller (and compact) group of *real* orthogonal matrices of a given size. (d) If $A \in M_n$ is orthogonal, show that $|\det A| = 1$ but that A could have eigenvalues λ with $|\lambda| \neq 1$. *Hint:* Consider $A(t)$ in (a) to show that $|\lambda(t)|$ can be arbitrarily large. (e) If $A \in M_n$

is orthogonal, show that $\bar{A}$, A^T, and A^* are all orthogonal and that A is nonsingular. Do the rows or columns of A form an orthogonal set? (f) Characterize the diagonal orthogonal matrices. Compare with Problem 4. To help avoid confusion, some authors refer to an orthogonal matrix which is not necessarily real as a *complex orthogonal matrix*, but this distinction is not always made in the literature; the term *orthogonal matrix* sometimes means what we have called a *real orthogonal matrix*.

9. If $U \in M_n$ is unitary, show that $\bar{U}$, U^T, and U^* are all unitary.

10. If $U \in M_n$ is unitary, show that $x, y \in \mathbf{C}^n$ are orthogonal if and only if Ux and Uy are orthogonal.

11. One might call a nonsingular matrix $A \in M_n$ *skew-orthogonal* if $A^{-1} = -A^T$. Show that A is skew-orthogonal if and only if $\pm iA$ is orthogonal. More generally, if $\theta \in \mathbf{R}$, show that $A^{-1} = e^{i\theta}A^T$ if and only if $e^{i\theta/2}A$ is orthogonal. What is this for $\theta = \pi$? for $\theta = 0$?

12. Show that if $A \in M_n$ is similar to a unitary matrix, then A^{-1} is similar to A^*.

13. Consider the matrix $\operatorname{diag}(2, \frac{1}{2}) \in M_2$ and show that the set of matrices that are similar to unitary matrices is a proper subset of the set of matrices A for which A^{-1} is similar to A^*.

14. Show that the intersection of the group of unitary matrices in M_n with the group of complex orthogonal matrices in M_n is the group of real orthogonal matrices in M_n. *Hint:* Consider $U = A + iB$, where $U, A, B \in M_n$ and A, B are real. If U is both unitary and complex orthogonal, show that $B^TB = 0$ and hence $(Be_i)^T(Be_i) = 0$ for each standard unit basis vector $e_i \in \mathbf{R}^n$, and hence every column of B is 0.

Further Reading. For more information about generalized unitary matrices that satisfy the conditions of Theorem (2.1.9), see C. R. DePrima and C. R. Johnson, "The Range of $A^{-1}A^*$ in $GL(n, \mathbf{C})$," *Linear Algebra Appl.* 9 (1974), 209–222.

2.2 Unitary equivalence

Since $U^* = U^{-1}$ for unitary U, the transformation on M_n given by $A \rightarrow U^*AU$ is a similarity transformation if U is unitary. This special type of similarity is called unitary similarity or unitary equivalence.

2.2.1 **Definition.** A matrix $B \in M_n$ is said to be *unitarily equivalent* to $A \in M_n$ if there is a unitary matrix $U \in M_n$ such that $B = U^*AU$. If U may

be taken to be real (and hence is real orthogonal), then B is said to be *(real) orthogonally equivalent* to A.

Exercise. Show that unitary equivalence is an equivalence relation.

2.2.2 **Theorem.** If $A = [a_{ij}]$ and $B = [b_{ij}] \in M_n$ are unitarily equivalent, then

$$\sum_{i,j=1}^{n} |b_{ij}|^2 = \sum_{i,j=1}^{n} |a_{ij}|^2$$

Proof: Observe that $\sum_{i,j} |a_{ij}|^2 = \operatorname{tr} A^*A$, by carrying out the matrix multiplication. Thus, it suffices to check that $\operatorname{tr} B^*B = \operatorname{tr} A^*A$. But if $B = U^*AU$, then $\operatorname{tr} B^*B = \operatorname{tr} U^*A^*UU^*AU = \operatorname{tr} U^*A^*AU = \operatorname{tr} A^*A$, because the trace is a similarity invariant. □

Exercise. Theorem (2.2.2) says that $\operatorname{tr} A^*A$ is a unitary similarity invariant. Carry out another proof without using A^*A, but using the fact from the preceding section that multiplication by a unitary matrix leaves the Euclidean length of a vector unchanged. Note that matrix multiplication from the left multiplies the columns, and matrix multiplication from the right multiplies the rows of a matrix.

Exercise. Show that

$$\begin{bmatrix} 3 & 1 \\ -2 & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$$

are similar but are not unitarily equivalent.

Since unitary equivalence implies similarity, but not conversely, the unitary equivalence relation partitions M_n into *finer* equivalence classes than the similarity equivalence relation. Unitary equivalence, like similarity, corresponds to a change of basis, but of a special type – a change from one *orthonormal* basis to another. An orthonormal change of basis leaves unchanged the sum of squares of the absolute values of the entries, a quantity that may be changed in a nonorthonormal change of basis. Unitary equivalence is computationally simpler than similarity because the conjugate transpose is much easier to compute than the inverse. It also better preserves accuracy in the presence of round-off errors, and is therefore preferable in numerical calculations. The precise reasons for this are not explained here, but an intuitive explanation lies in the length-preserving nature of multiplication by a unitary matrix.

Two special (and very simple) types of unitary matrices give unitary

equivalence transformations that are very important in eigenvalue calculations.

2.2.3 Example: plane rotations. Let $U(\theta; i, j)$

$$
= \left[\begin{array}{ccccccc}
1 & & & & & & \\
\ddots & & & 0 & & & \\
1 & & & & & & \\
\text{--------} & \cos\theta & 0 \;\cdots & 0 & -\sin\theta & \text{--------} & \\
& 0 & 1 & & & & \\
0 & \vdots & & \ddots & & 0 & \\
& 0 & 0 & 1 & & & \\
\text{--------} & \sin\theta & 0 \;\cdots & 0 & \cos\theta & \text{--------} & \\
& \vdots & & & & 1 & \\
& & & 0 & & \ddots & \\
& & & & & & 1
\end{array}\right]
\begin{array}{l}
\\ \\ \\ \text{---- row } i \\ \\ \\ \\ \text{---- row } j \\ \\ \\ \\
\end{array}
$$

$$\text{column } i \qquad\qquad \text{column } j$$

This is simply the identity matrix, with the i, i and j, j entries replaced by $\cos\theta$ and the i, j entry (respectively j, i entry) replaced by $-\sin\theta$ (respectively $\sin\theta$).

Exercise. Verify that $U(\theta; i, j)$ is an orthogonal matrix in $M_n(\mathbf{R})$ for any pair of indices $1 \le i < j \le n$ and any angle parameter $0 \le \theta < 2\pi$. The matrix $U(\theta; i, j)$ simply carries out a rotation (through an angle θ) in the i, j coordinate plane. Notice that left multiplication by $U(\theta; i, j)$ affects only rows i and j and right multiplication by $U(\theta; i, j)$ affects only columns i and j of the matrix multiplied. Thus, under unitary equivalence via $U(\theta; i, j)$, only rows and columns i and j are changed. Unitary equivalences via plane rotations are the basic feature of eigenvalue calculation schemes of Jacobi and Givens (see Problems 1 and 2).

2.2.4 Example: Householder transformations. Let $w \in \mathbf{C}^n$ be a nonzero vector and define $U_w \in M_n$ by $U_w = I - tww^*$ in which $t = 2(w^*w)^{-1}$. Note that $ww^* \in M_n$ and w^*w is a positive scalar. If w were normalized $(w^*w = 1)$, t would be 2 and U_w would be $I - 2ww^*$. Often the matrix U_w is presented assuming that w is already normalized.

Exercise. Show that U_w acts as the identity on the complementary subspace $w^\perp$ and that it acts as a reflection on the one-dimensional subspace spanned by w; that is, $U_w x = x$ if $x \perp w$ and $U_w w = -w$.

Exercise. Show that U_w is both unitary and Hermitian $(U_w^* = U_w)$. A matrix of the form U_w is called a *Householder transformation*. Unitary equivalence via a U_w is sometimes also referred to as a Householder transformation. These transformations arise in a number of contexts including the eigenvalue calculation scheme of Householder (see Problems 4 and 5) and other unitary reductions. Note that Householder transformations in general change all the entries of a matrix or vector to which they are applied, but they provide extremely efficient and accurate reductions for a number of uses.

Theorem (2.2.2) provides a necessary but not sufficient condition for two given matrices to be unitarily equivalent. It can be augmented with additional identities that collectively do provide necessary and sufficient conditions. A key role is played by the following simple notion. Let s, t be two given noncommuting variables. We refer to any finite formal product of nonnegative powers of s, t

$$W(s,t) = s^{m_1}t^{n_1}s^{m_2}t^{n_2}\cdots s^{m_k}t^{n_k}, \qquad m_1, n_1, \ldots, m_k, n_k \ge 0 \qquad (2.2.5)$$

as a *word in s and t*. The *degree* of the word $W(s,t)$ is the nonnegative integer $m_1 + n_1 + m_2 + n_2 + \cdots + m_k + n_k$, that is, the sum of all the exponents in the word. If $A \in M_n$ is given, we may formally define a *word in A and A^** as

$$W(A, A^*) = A^{m_1}(A^*)^{n_1}A^{m_2}(A^*)^{n_2}\cdots A^{m_k}(A^*)^{n_k}$$

Since the powers of A and A^* need not commute, it may not be possible to simplify the expression of $W(A, A^*)$ by rearranging the terms in the product.

If A is unitarily equivalent to some $B \in M_n$, then $A = UBU^*$ for some unitary $U \in M_n$ and one computes readily that

$$\begin{aligned}
W(A,A^*) &= (UBU^*)^{m_1}(UB^*U^*)^{n_1}\cdots(UBU^*)^{m_k}(UB^*U^*)^{n_k}\\
&= UB^{m_1}U^*U(B^*)^{n_1}U^*\cdots UB^{m_k}U^*U(B^*)^{n_k}U^*\\
&= UB^{m_1}(B^*)^{n_1}\cdots B^{m_k}(B^*)^{n_k}U^*\\
&= UW(B,B^*)U^*
\end{aligned}$$

Thus, $\operatorname{tr} W(A, A^*) = \operatorname{tr} UW(B, B^*)U^* = \operatorname{tr} W(B, B^*)$. If we take the word $W(s,t) = ts$, we obtain the identity in Theorem (2.2.2).

If one considers all possible words $W(s,t)$, this observation gives infinitely many necessary conditions for two matrices to be unitarily equivalent. A theorem of W. Specht, which we state without proof, guarantees that these infinitely many necessary conditions are also sufficient.

2.2.6 Theorem. Two given matrices $A, B \in M_n$ are unitarily equivalent if and only if

$$\operatorname{tr} W(A, A^*) = \operatorname{tr} W(B, B^*) \tag{2.2.7}$$

for every word $W(s, t)$ in two noncommuting variables.

Specht's theorem can be used to show that two given matrices are *not* unitarily equivalent, but except in special situations (see Problem 6), it may be useless in showing that two given matrices *are* unitarily equivalent because infinitely many conditions must be verified. Fortunately, there is an improvement to Specht's theorem due to C. Pearcy that says it suffices to check the trace identities (2.2.7) for only finitely many words.

2.2.8 Theorem. Two given matrices $A, B \in M_n$ are unitarily equivalent if and only if $\operatorname{tr} W(A, A^*) = \operatorname{tr} W(B, B^*)$ for every word $W(s, t)$ of degree at most $2n^2$.

The finite bound in Pearcy's theorem is a vast improvement on Specht's theorem, but even it is known to be extremely conservative. For $n = 2$, it actually suffices to check the trace identities (2.2.7) for the three words $W(s, t) = s$, s^2, and ts rather than consider all the words of degree at most $2(2^2) = 8$. For $n = 3$, it suffices to check the trace identities for the nine words $W(s, t) = s$, s^2, ts, s^3, ts^2, t^2s^2, $tsts$, ts^2ts, and ts^2t^2s rather than consider all the words of degree at most $2(3^2) = 18$.

Problems

1. Let $A = [a_{ij}] \in M_n(\mathbf{R})$ be symmetric ($A^T = A$), but *not* diagonal, and suppose that indices $i \neq j$ are chosen so that $|a_{ij}|$ is as large as possible. Define θ by $(a_{ii} - a_{jj})/2a_{ij} = \cot(2\theta)$ and let $U(\theta; i, j)$, as given in (2.2.3), be the resulting plane rotation. Use (2.2.2) to show that if $B = U(\theta; i, j)^* A U(\theta; i, j) = (b_{ij})$, then $\sum_{i \neq j} |b_{ij}|^2 < \sum_{i \neq j} |a_{ij}|^2$. Show that repeated applications of such plane rotations (chosen in the same way for B and its successors) will decrease the sums of the squares of the *off-diagonal* entries while preserving the sums of the squares of *all* the entries; at each step, the matrix is "more nearly diagonal" than at the step before. This is the method of Jacobi for calculating the eigenvalues of a real symmetric matrix. This method produces a sequence of matrices that converges to a real diagonal matrix. Why must the diagonal entries of the limit be the eigenvalues of A?

2. The eigenvalue calculation method of Givens for real symmetric matrices (or general real matrices) also utilizes plane rotations, but in a rather different way. Show that a symmetric matrix $A = [a_{ij}] \in M_n(\mathbf{R})$ is orthogonally equivalent to a tridiagonal (symmetric) matrix via plane rotations and that a general $A \in M_n(\mathbf{R})$ is orthogonally equivalent to (lower) Hessenberg form via plane rotations. See (0.9.9) and (0.9.10) for tridiagonal and Hessenberg matrices. *Hint:* Choose a plane rotation $U_{1,3}$ of the form $U(\theta; 2, 3)$ in which (2.2.9b) is used to choose θ so that the 1, 3 entry of $U_{1,3}^* A U_{1,3}$ is 0. Choose another plane rotation of the form $U(\theta; 2, 4)$ and continue in this way to zero out the rest of the first row. Then start on the second row beginning with the 2, 4 entry. Choose a plane rotation of the form $U(\theta; 3, 4)$ to zero out the 2, 4 entry and so on. Note that this process does not disturb previously manufactured 0 entries. Note also that orthogonal equivalence preserves symmetry. The characteristic polynomial of a tridiagonal matrix may be determined easily and the eigenvalues determined by a root-finding scheme. Notice that Givens's method produces a tridiagonal matrix after finitely many plane rotations, but it does not display the eigenvalues or the eigenvectors, which must then be obtained from some further calculation. Jacobi's method does not, in general, terminate after finitely many plane rotations, but it tries to produce a diagonal matrix as well as an orthonormal set of eigenvectors.

3. Show that every matrix $A \in M_n$ is unitarily equivalent to a matrix with equal main diagonal entries. *Hint:* (a) If $A \in M_2$, consider $A - (1/2)(\operatorname{tr} A)I$ to show that it suffices to consider only the case $\operatorname{tr} A = 0$. If $x \in \mathbf{C}^2$ is a unit vector such that $x^*Ax = 0$, let $U = [x\ y] \in M_2$ be unitary, show that the 1, 1 entry of U^*AU is zero, and use the trace condition to show that the 2, 2 entry is zero as well. To find such a vector x, let w, z be unit eigenvectors associated with the two eigenvalues $\pm\lambda$ of A. If $\lambda = 0$ take $x = w$. If $\lambda \neq 0$ let $x(\theta) \equiv e^{i\theta}w + z$; show that $x(\theta) \neq 0$ for all $\theta \in \mathbf{R}$, and $x(\theta)^*Ax(\theta) = 0$ for some $\theta \in \mathbf{R}$; let $x = x(\theta)/[x(\theta)^*x(\theta)]^{1/2}$ for this θ. *Remark:* If $A \in M_2(\mathbf{R})$, it is easy to construct a (real) plane rotation $U = U(\theta; 1, 2)$ that makes the diagonal entries of U^TAU equal, but this does not help in the complex case. (b) For $A = [a_{ij}] \in M_n$, define $f(A) \equiv \max\{|a_{ii} - a_{jj}| : i, j = 1, 2, \ldots, n\}$ and let $A_2 \equiv \left[\begin{smallmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{smallmatrix}\right]$ for a pair of indices i, j for which $f(A) = |a_{ii} - a_{jj}|$. Let $U_2 \in M_2$ be a unitary matrix such that $U_2^* A_2 U_2$ has equal main diagonal entries. Construct $U(i, j) \in M_n$ from U_2 in the same way that $U(\theta; i, j)$ was constructed from a 2-by-2 plane rotation in (2.2.3) and show that $f(U(i, j)^* A U(i, j)) < f(A)$ if there are not ties for the maximum absolute diagonal difference; if there are ties, this construction can be repeated. Conclude that if $f(A) \neq 0$, there is a unitary $U \in M_n$ such that $f(U^*AU) < f(A)$. Let $R(A) \equiv \{U^*AU : U \in M_n \text{ is unitary}\}$. Show that $R(A)$ is compact and note that f is a continuous function on $R(A)$. Let $C \in R(A)$ be such that $f(C) = \min\{f(B) : B \in R(A)\}$. Show that $f(C) > 0$ is impossible and that the assertion follows from $f(C) = 0$.

4. Show that any vector $x \in \mathbf{R}^n$ with Euclidean length $r = (x^Tx)^{1/2}$ may be transformed into any other vector $y \in \mathbf{R}^n$ of length r, $y \neq x$, by a Householder transformation, that is, $y = U_w x$ for some $w \in \mathbf{R}^n$. *Hint:* Try $w = x - y$. What can you say if $x, y \in \mathbf{C}^n$?

5. Householder's method for calculating eigenvalues of $A \in M_n(\mathbf{R})$, like the method of Givens, first reduces A to (upper) Hessenberg form (or tridiagonal form in the symmetric case). Show constructively that a matrix of the form

$$\left[\begin{array}{cccccc|c} * & * & * & \cdots & * & & \\ * & * & * & & & & \\ & * & * & & \vdots & & \\ & & * & \ddots & & & * \\ & 0 & & \ddots & & & \\ & & & & * & & \\ & & & & * & & \\ \hline & & 0 & & & & * \end{array}\right] \begin{array}{l} \\ \\ \\ \\ \\ \leftarrow \text{row } k \\ \leftarrow \text{row } k+1 \\ \\ \end{array}$$

$\uparrow$
column k

with 0's below the $(j+1)$st entry in the jth column, $j=1,\ldots,k$, may be transformed to a matrix of the same form, $j=1,\ldots,k+1$, via a single real orthogonal similarity using a Householder transformation. Conclude that any matrix $A \in M_n(\mathbf{R})$ may be reduced to upper Hessenberg form by a sequence of $(n-2)$ Householder similarities and that a symmetric matrix $A \in M_n(\mathbf{R})$ may be reduced to tridiagonal form in the same way. *Hint:* For the $(k+1)$st column, choose a Householder transformation $U \in M_{n-k-1}$ that transforms the $(n-k-1)$ vector of entries occurring below the diagonal to an appropriate multiple of $[1,0,\ldots,0]^T \in \mathbf{R}^{n-k-1}$. Then transform the entire matrix by a similarity via the orthogonal matrix

$$\begin{bmatrix} I & 0 \\ 0 & U \end{bmatrix} \in M_n$$

and see that the desired 0 pattern prevails.

6. Let $A \in M_n$ and $B, C \in M_m$ be given. Use either Specht's theorem (2.2.6) or Pearcy's theorem (2.2.8) to show that B and C are unitarily equivalent if and only if any one of the following conditions holds:

(a) $\left[\begin{smallmatrix} A & 0 \\ 0 & B \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} A & 0 \\ 0 & C \end{smallmatrix}\right]$ are unitarily equivalent.

(b) $\begin{bmatrix} B & & 0 \\ & B & \\ 0 & \ddots & B \end{bmatrix}$ and $\begin{bmatrix} C & & 0 \\ & C & \\ 0 & \ddots & C \end{bmatrix}$ are unitarily equivalent, where both direct sums contain the same number of terms.

(c) $\begin{bmatrix} A & & 0 \\ & B & \\ 0 & \ddots & B \end{bmatrix}$ and $\begin{bmatrix} A & & 0 \\ & C & \\ 0 & \ddots & C \end{bmatrix}$ are unitarily equivalent, where both direct sums contain the same number of terms.

7. Show that there are 2^{k-1} distinct words $W(s,t)$ of the form (2.2.5) that have a given degree k and conclude that there are at most 4^{n^2} distinct words of degree at most $2n^2$.

8. Give an example of two 2-by-2 matrices that satisfy the identity (2.2.2) but are not unitarily equivalent. Explain why.

Further Readings and Notes. For the original proof of Theorem (2.2.6), see W. Specht, "Zur Theorie der Matrizen II," *Jahresbericht der Deutschen Mathematiker Vereinigung* 50 (1940), 19–23. Theorem (2.2.8) is proved in C. Pearcy, "A Complete Set of Unitary Invariants for Operators Generating Finite W^*-Algebras of Type I," *Pacific J. Math.* 12 (1962), 1405–1416. A discussion of unitary equivalence of low-order matrices is in C. Pearcy, "A Complete Set of Unitary Invariants for 3×3 Complex Matrices," *Trans. Amer. Math. Soc.* 104 (1962), 425–429.

2.3 Schur's unitary triangularization theorem

Perhaps the most fundamentally useful fact of elementary matrix theory is that any matrix $A \in M_n$ is unitarily equivalent to an upper triangular matrix T [and also to a lower triangular matrix]. The diagonal entries of T are, of course, the eigenvalues of A. Although this form is far from unique, it represents the simplest form achievable under unitary equivalence.

2.3.1 **Theorem** (Schur). Given $A \in M_n$ with eigenvalues $\lambda_1,\ldots,\lambda_n$ in any prescribed order, there is a unitary matrix $U \in M_n$ such that

$$U^*AU = T = [t_{ij}]$$

is upper triangular, with diagonal entries $t_{ii} = \lambda_i$, $i=1,\ldots,n$. That is, every square matrix A is unitarily equivalent to a triangular matrix whose diagonal entries are the eigenvalues of A in a prescribed order. Furthermore, if $A \in M_n(\mathbf{R})$ and if all the eigenvalues of A are real, then U may be chosen to be real and orthogonal.

Proof: The proof is algorithmic and proceeds by a sequence of reductions of like type. Let $x^{(1)}$ be a normalized eigenvector of A associated with the eigenvalue λ_1. The nonzero vector $x^{(1)}$ may be extended to a basis

$$x^{(1)}, y^{(2)}, y^{(3)}, \ldots, y^{(n)}$$

of $\mathbf{C}^n$. Apply the Gram-Schmidt orthonormalization procedure (0.6.4) to this basis to produce an orthonormal basis

$$x^{(1)}, z^{(2)}, \ldots, z^{(n)}$$

of $\mathbf{C}^n$. Array these orthonormal vectors left to right as the columns of a unitary matrix U_1. Since the first column of AU_1 is $\lambda_1 x^{(1)}$, a calculation reveals that $U_1^*(AU_1)$ has the form

$$U_1^*AU_1 = \left[\begin{array}{c|c} \lambda_1 & * \\ \hline 0 & A_1 \end{array}\right]$$

The matrix $A_1 \in M_{n-1}$ has eigenvalues $\lambda_2, \ldots, \lambda_n$. Let $x^{(2)} \in \mathbf{C}^{n-1}$ be a normalized eigenvector of A_1 corresponding to λ_2, and do it all over again. Determine a unitary $U_2 \in M_{n-1}$ such that

$$U_2^*A_1U_2 = \left[\begin{array}{c|c} \lambda_2 & * \\ \hline 0 & A_2 \end{array}\right]$$

and let

$$V_2 = \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & U_2 \end{array}\right]$$

The matrices V_2 and U_1V_2 are then unitary, and $V_2^*U_1^*AU_1V_2$ has the form

$$V_2^*U_1^*AU_1V_2 = \left[\begin{array}{cc|c} \lambda_1 & * & \\ 0 & \lambda_2 & * \\ \hline \multicolumn{2}{c|}{0} & A_2 \end{array}\right]$$

Continue this reduction to produce unitary matrices $U_i \in M_{n-i+1}$, $i = 1, \ldots, n-1$ and unitary matrices $V_i \in M_n$, $i = 2, \ldots, n-1$. The matrix

$$U = U_1V_2V_3 \cdots V_{n-1}$$

is unitary and U^*AU yields the desired form.

If all eigenvalues of $A \in M_n(\mathbf{R})$ happen to be real, then the corresponding eigenvectors can be chosen to be real and all the above steps may be carried out in real arithmetic, verifying the final assertion. □

Remark: Follow the proof of (2.3.1) to see that "upper triangular" could be replaced by "lower triangular" in the statement of the theorem with, of course, a different unitary equivalence U.

2.3.2 **Example.** Neither the unitary matrix U nor the triangular matrix T of Theorem (2.3.1) is unique. Not only may the diagonal entries of T

(the eigenvalues of A) appear in any order, but unitarily equivalent upper triangular matrices may appear very different above the diagonal. For example,

$$T_1 = \begin{bmatrix} 1 & 1 & 4 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{bmatrix} \quad \text{and} \quad T_2 = \begin{bmatrix} 2 & -1 & 3\sqrt{2} \\ 0 & 1 & \sqrt{2} \\ 0 & 0 & 3 \end{bmatrix}$$

are unitarily equivalent via

$$U = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & \sqrt{2} \end{bmatrix}$$

In general, many different upper triangular matrices can be in the same unitary equivalence class.

Remark: Notice that the technique of the proof (2.3.1) is simply that of sequential deflation, as outlined in Problem 8 in Section (1.4).

Exercise. If $A \in M_n$ is unitarily equivalent to an upper triangular matrix $T = [t_{ij}] \in M_n$, the entries t_{ij} are not uniquely determined, but the quantity $\sum_{i<j} |t_{ij}|^2$ is uniquely determined. Determine the value of $\sum_{i<j} |t_{ij}|^2$ in terms of the entries and eigenvalues of A. *Hint:* Use (2.2.2).

Exercise. If $A = [a_{ij}]$ and $B = [b_{ij}] \in M_2$ are similar and if $\sum_{i,j} |a_{ij}|^2 = \sum_{i,j} |b_{ij}|^2$, show that A and B are unitarily equivalent. Show by example that this is not the case in higher dimensions. *Hint:* Notice that if A and B are unitarily equivalent, then so are $A + A^*$ and $B + B^*$. Consider

$$A = \begin{bmatrix} 1 & 3 & 0 \\ 0 & 2 & 4 \\ 0 & 0 & 3 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 5 \\ 0 & 0 & 3 \end{bmatrix}$$

It is a useful adjunct to (2.3.1) that a commuting family of matrices may be simultaneously upper triangularized.

2.3.3 **Theorem.** Let $\mathfrak{F} \subseteq M_n$ be a commuting family. There is a unitary matrix $U \in M_n$ such that U^*AU is upper triangular for every $A \in \mathfrak{F}$.

Proof: Return to the proof of (2.3.1). Exploiting (1.3.17) at each step of the proof in which a choice of an eigenvector (and unitary matrix) is made, the *same* eigenvector (and unitary matrix) may be chosen for every $A \in \mathfrak{F}$. Moreover, unitary equivalence preserves commutativity,

and a partitioned multiplication calculation reveals that, if two matrices of the form

$$\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix}$$

commute, then A_{22} and B_{22} commute also. Thus, the commuting family property is inherited by each A_i at each stage in the reduction process of the proof of (2.3.1). We conclude that all ingredients in the U of (2.3.1) may be chosen in the same way for all members of a commuting family, thus verifying (2.3.3). Notice that we do not claim that any special order may be chosen for the eigenvalues of the various family members. We simply take them as they come using (1.3.17). □

A strictly real version of (2.3.1) is contained in the following theorem.

2.3.4 **Theorem.** If $A \in M_n(\mathbf{R})$, there is a real orthogonal matrix $Q \in M_n(\mathbf{R})$ such that

$$Q^T A Q = \begin{bmatrix} A_1 & & * \\ & A_2 & \\ 0 & & \ddots \\ & & & A_k \end{bmatrix} \in M_n(\mathbf{R}), \qquad 1 \le k \le n \tag{2.3.5}$$

where each A_i is a real 1-by-1 matrix, or a real 2-by-2 matrix with a nonreal pair of complex conjugate eigenvalues. The diagonal blocks A_i may be arranged in any prescribed order.

One cannot, in general, hope to reduce a real matrix to upper triangular form by a real similarity (let alone a real orthogonal similarity) because the diagonal entries would then be eigenvalues, which could be nonreal. The form (2.3.5) is the most nearly triangular form one can achieve by a real orthogonal similarity. It will not be upper triangular if A has any nonreal eigenvalues, but it will always be in upper Hessenberg form.

Exercise. Modify the argument for (2.3.1) to prove (2.3.4). *Hint:* If λ is a real eigenvalue of the real matrix A, then there is a corresponding real eigenvector that can be used to deflate A as in (2.3.1). If $\lambda = \alpha + i\beta$ is a nonreal eigenvalue for A and if $Ax = \lambda x$, $x = u + iv \neq 0$, $u, v \in \mathbf{R}^n$, show that $Au = \alpha u - \beta v$, $Av = \alpha v + \beta u$, and $A\bar{x} = \bar{\lambda}\bar{x}$, and that $\{x, \bar{x}\}$ is an independent set. Deduce that $\{u, v\}$ is an independent set and apply the Gram–Schmidt procedure to it to obtain a real orthonormal set $\{w, z\}$.

Let Q_1 be a real orthogonal matrix whose first two columns are w and z. Show that

$$Q_1^T A Q_1 = \left[\begin{array}{cc|c} * & * & * \\ * & * & \\ \hline 0 & & \tilde{A} \end{array}\right]$$

so that A may be deflated two columns at a time in this case. Notice that blocks A_i corresponding to each real eigenvalue and each pair of complex conjugate eigenvalues can be arranged in any prescribed order in (2.3.5).

There is also a real version of (2.3.3).

2.3.6 **Theorem.** Let $\mathfrak{F} \subseteq M_n(\mathbf{R})$ be a commuting family. There is a real orthogonal matrix $Q \in M_n(\mathbf{R})$ such that $Q^T A Q$ is of the form (2.3.5) for every $A \in \mathfrak{F}$.

Exercise. Modify the proof of (2.3.3) to prove (2.3.6). *Hint:* First deflate all members of $\mathfrak{F}$ using all the common real eigenvectors. Then consider the common nonreal eigenvectors and deflate two columns at a time as in the proof of (2.3.4). Notice that different members of $\mathfrak{F}$ may have different numbers of 2-by-2 diagonal blocks after the common real orthogonal similarity, but if one member has a 2-by-2 block in a certain position and another member does not, then the latter must have a pair of equal 1-by-1 blocks there.

Problems

1. Let $x \in \mathbf{C}^n$ be a given unit vector ($x^*x = 1$) and write $x = [x_1\ y^T]^T$, where $x_1 \in \mathbf{C}$ and $y \in \mathbf{C}^{n-1}$. Choose $\theta \in \mathbf{R}$ such that $e^{i\theta}x_1 \ge 0$ and define $z = e^{i\theta}x = [z_1\ \zeta^T]^T$, where $z_1 \in \mathbf{R}$ is nonnegative and $\zeta \in \mathbf{C}^{n-1}$. Show that the matrix

$$V = \left[\begin{array}{c|c} z_1 & \zeta^* \\ \hline \zeta & -I + \dfrac{1}{1+z_1}\zeta\zeta^* \end{array}\right]$$

is unitary. *Hint:* Compute $V^*V = V^2$. Conclude that the matrix $U = e^{-i\theta}V = [x\ u_2\ \ldots\ u_n]$ is a unitary matrix whose first column is the given vector x. This gives a constructive method to obtain the unitary matrices needed for the successive deflation steps in the proof of Schur's theorem (2.3.1).

2. If $x \in \mathbf{R}^n$ is a given unit vector, show how to streamline the construction described in Problem 1 to produce a real orthogonal matrix $Q \in M_n(\mathbf{R})$ whose first column is x. Prove that your construction works.

3. Let $A \in M_n(\mathbf{R})$. Explain why the nonreal eigenvalues of A (if any) must occur in conjugate pairs.

4. Consider the family

$$\mathfrak{F} = \left\{ \begin{bmatrix} 0 & -1 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix} \right\}$$

and show that the hypothesis of commutativity in Theorem (2.3.3), while sufficient to imply simultaneous unitary upper triangularizability of $\mathfrak{F}$, is not necessary.

5. Let $\mathfrak{F} = \{A_1, \ldots, A_k\} \subset M_n$ be a given family, and let

$$\mathcal{G} = \{A_i A_j : i, j = 1, 2, \ldots, k\}$$

be the family of all pair-wise products of matrices in $\mathfrak{F}$. It is a fact that if $\mathcal{G}$ is commutative, then $\mathfrak{F}$ can be simultaneously unitarily upper triangularized if and only if every eigenvalue of every commutator $A_i A_j - A_j A_i$ is zero. Show that assuming commutativity of $\mathcal{G}$ is a weaker hypothesis than assuming commutativity of $\mathfrak{F}$. Show that the family $\mathfrak{F}$ in Problem 4 has a corresponding $\mathcal{G}$ that is commutative, and that it also satisfies the zero eigenvalue condition.

6. Let $A, B \in M_n$ be given, and suppose A and B are simultaneously similar to upper triangular matrices; that is, $S^{-1}AS$ and $S^{-1}BS$ are both upper triangular for some nonsingular $S \in M_n$. Show that every eigenvalue of $AB - BA$ must be zero. *Hint:* If $\Delta_1, \Delta_2 \in M_n$ are both upper triangular, what is the main diagonal of $\Delta_1 \Delta_2 - \Delta_2 \Delta_1$?

7. Although every square matrix can be reduced to upper triangular form by unitary similarity, this is not true for complex orthogonal similarity. If a given $A \in M_n$ can be written as $A = Q \Delta Q^T$, where $Q \in M_n$ is complex orthogonal and $\Delta \in M_n$ is upper triangular, show that A has at least one eigenvector $x \in \mathbf{C}^n$ such that $x^T x \neq 0$. Consider $A = \left[\begin{smallmatrix} 1 & i \\ i & -1 \end{smallmatrix}\right]$ to show that not every $A \in M_n$ can be upper-triangularized by a complex orthogonal similarity.

8. Let $Q \in M_n$ be a given complex orthogonal matrix, and suppose $x \in \mathbf{C}^n$ is an eigenvector of Q associated with an eigenvalue $\lambda \neq \pm 1$. Show that $x^T x = 0$. *Hint:* Multiply both sides of the identity $Qx = \lambda x$ by its transpose. See Problem 8(a) of Section (2.1) for an example of a family of 2-by-2 complex orthogonal matrices with both eigenvalues different from ± 1. Show that none of these matrices can be reduced to upper triangular form by orthogonal similarity.

Further Reading. For a proof of the stronger form of Theorem (2.3.3) asserted in Problem 5, see Y. P. Hong and R. A. Horn, "On Simultaneous Reduction of Families of Matrices to Triangular or Diagonal Form by Unitary Congruences," *Linear and Multilinear Algebra* 17 (1985), 271–288.

2.4 Some implications of Schur's theorem

Several elementary consequences of Schur's unitary triangularization illustrate its utility.

Exercise. Use (2.3.1) to show that if $A \in M_n$ has eigenvalues $\lambda_1, \ldots, \lambda_n$, counting multiplicity, then $\det A = \prod_{i=1}^n \lambda_i$ and $\operatorname{tr} A = \sum_{i=1}^n \lambda_i$. Recall that this was proved in another way in Chapter 1. *Hint:* For the trace, recall that $\operatorname{tr} AB = \operatorname{tr} BA$ follows from a direct calculation, so that the trace is similarity invariant. What about the other elementary symmetric functions of the eigenvalues?

The fact that every matrix satisfies its own characteristic equation (2.4.2) follows from Schur's theorem and a simple observation about multiplication of triangular matrices.

2.4.1 **Lemma.** Suppose that $R = [r_{ij}]$ and $T = [t_{ij}] \in M_n$ are upper triangular and that $r_{ij} = 0$, $1 \le i, j \le k < n$, and $t_{k+1,k+1} = 0$. Let $T' = [t'_{ij}] = RT$. Then $t'_{ij} = 0$, $1 \le i, j \le k+1$.

Proof: Since $R(\{1, 2, \ldots, k\}) = 0$ and $t_{k+1,k+1} = 0$, R and T have the form

$$R = \left[\begin{array}{c|c} 0 & * \\ \hline 0 & \begin{matrix} * & & * \\ & \ddots & \\ 0 & & * \end{matrix} \end{array}\right], \qquad T = \left[\begin{array}{c|c} \begin{matrix} * & & * \\ & \ddots & \\ 0 & & * \end{matrix} & * \\ \hline 0 & \begin{matrix} 0 & & * \\ & * & \\ & & \ddots \\ 0 & & * \end{matrix} \end{array}\right]$$

where both upper left blocks in the partitions are k-by-k. The upper left k-by-k block of T' is clearly 0 by partitioned multiplication [see (0.7) for

notation and elementary facts]. Also, inspection reveals that the first $k+1$ rows of R have 0's in all nonzero positions of column $k+1$ of T, and that the first $k+1$ columns of T have 0's in all nonzero positions of row $k+1$ of R. Matrix multiplication then shows that T' (partitioned in the same way) has the form

$$T' = \left[\begin{array}{c|ccc} \mathbf{0} & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} & & * \\ \hline \mathbf{0} & \begin{matrix} 0 \\ \vdots \\ \vdots \\ 0 \end{matrix} & \begin{matrix} * & & \\ & \ddots & \\ 0 & & * \end{matrix} & * \end{array}\right]$$

and $T'(\{1, \ldots, k+1\}) = 0$, as was to be shown. □

Exercise. Show that the product of two upper triangular matrices is upper triangular, and show that the product of two similarly partitioned block upper triangular matrices is block upper triangular.

Exercise. Generalize (2.4.1) by showing that if R and T are upper triangular and $T' = RT$, then

$$T'(\{i, i+1, \ldots, i+j\}) = R(\{i, i+1, \ldots, i+j\})T(\{i, i+1, \ldots, i+j\})$$

2.4.2 **Theorem** (Cayley–Hamilton). Let $p_A(t)$ be the characteristic polynomial of $A \in M_n$. Then

$$p_A(A) = 0$$

Proof: Since $p_A(t)$ is of degree n with leading coefficient 1 and the roots of $p_A(t) = 0$ are precisely the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A, counting multiplicity, we may factor $p_A(t)$ as

$$p_A(t) = (t-\lambda_1)(t-\lambda_2)\cdots(t-\lambda_n)$$

Using (2.3.1), write A as

$$A = UTU^*$$

where T is upper triangular with λ_i in the ith diagonal position, $i = 1, \ldots, n$. Now compute

$$\begin{aligned} p_A(A) &= p_A(UTU^*) = (UTU^* - \lambda_1 I)(UTU^* - \lambda_2 I)\cdots(UTU^* - \lambda_n I) \\ &= [U(T-\lambda_1 I)U^*][U(T-\lambda_2 I)U^*]\cdots[U(T-\lambda_n I)U^*] \end{aligned}$$

$$\begin{aligned} &= U[(T-\lambda_1 I)(T-\lambda_2 I)\cdots(T-\lambda_n I)]U^* \\ &= U p_A(T) U^* \end{aligned}$$

and notice that $p_A(A) = 0$ if and only if $p_A(T) = 0$. However, Lemma (2.4.1) allows us to conclude that $p_A(T) = 0$. The upper left 1-by-1 block of $T - \lambda_1 I$ is 0, and the 2, 2 entry of $T - \lambda_2 I$ is 0; since both are upper triangular, the upper left 2-by-2 block of $(T-\lambda_1 I)(T-\lambda_2 I)$ is 0. Inductively, since the upper left k-by-k block of $(T-\lambda_1 I)\cdots(T-\lambda_k I)$ and the $k+1, k+1$ entry of $(T-\lambda_{k+1} I)$ are 0 and both are upper triangular, the upper left $(k+1)$-by-$(k+1)$ block of $(T-\lambda_1 I)\cdots(T-\lambda_{k+1} I)$ is 0. Continuation until n is reached allows us to conclude that the product $p_A(T) = (T-\lambda_1 I)\cdots(T-\lambda_n I) = 0$, which completes the proof. □

Exercise. What is wrong with the following argument for the statement $p_A(A) = 0$? "Since $p_A(\lambda) = 0$ for every eigenvalue λ of $A \in M_n$, and since the eigenvalues of $q(A)$, q a polynomial, are the $q(\lambda)$, it follows that all eigenvalues of $p_A(A)$ are 0. Therefore, $p_A(A)$ is 0." This is a common mistaken argument for the Cayley–Hamilton theorem. Give an explicit example to illustrate where it is in error.

Exercise. What is wrong with this argument? "Since $p_A(t) = \det(tI - A)$, $p_A(A) = \det(AI - A) = \det(A - A) = \det 0 = 0$. Therefore, $p_A(A) = 0$."

If $p_A(t) = \det(tI - A)$ denotes the *characteristic polynomial* of $A \in M_n$, the *characteristic equation* is $p_A(t) = 0$. The roots of the characteristic equation are the eigenvalues of A. The Cayley–Hamilton theorem is often paraphrased as "every square matrix satisfies its own characteristic equation," but this must be understood carefully: The scalar polynomial $p_A(t)$ is first computed as $p_A(t) = \det(tI - A)$, and one then forms the matrix $p_A(A)$ from the characteristic polynomial.

We have proved the Cayley–Hamilton theorem for matrices with complex entries, and hence it must hold for matrices whose entries come from any subfield of the complex numbers (the reals or the rationals, for example). In fact, the Cayley–Hamilton theorem is a completely formal result that holds for matrices whose entries come from any field or, more generally, any commutative ring. See Problem 3.

One important use of the Cayley–Hamilton theorem is to write powers A^k of $A \in M_n$, for $k \geq n$, as linear combinations of $I, A, A^2, \ldots, A^{n-1}$. By a linear dependence argument, it is easy to show (since the dimension of M_n, considered as a vector space over the complex numbers, is n^2) that powers A^{n^2} and beyond can be expressed as linear combinations of lower powers, but the Cayley–Hamilton theorem provides a notable improvement.

2.4.3 Example. Let

$$A=\begin{bmatrix} 3 & 1 \\ -2 & 0 \end{bmatrix}$$

Then $p_A(t)=t^2-3t+2$, and $A^2-3A+2I=0$. Thus, $A^2=3A-2I$; $A^3=A(A^2)=3A^2-2A=3(3A-2I)-2A=7A-6I$; $A^4=7A^2-6A=15A-14I$, and so on. Also, since the constant term in $p_A(t)$, the determinant of A, is nonzero, A is nonsingular, and we may write A^{-1} as a polynomial in A. Again from $p_A(A)=A^2-3A+2I=0$, we get $2I=-A^2+3A=A(-A+3I)$, or

$$I=A[\tfrac{1}{2}(-A+3I)]$$

This means that $A^{-1}=-\frac{1}{2}A+\frac{3}{2}I=\begin{bmatrix} 0 & -1/2 \\ 1 & 3/2 \end{bmatrix}$.

Exercise. Given $A\in M_n$ with characteristic polynomial

$$p_A(t)=t^n+a_{n-1}t^{n-1}+a_{n-2}t^{n-2}+\cdots+a_1t+a_0$$

write A^n as a polynomial, of degree at most $n-1$, in A. Do the same for the next few powers of A after $n-1$. Assume further that A is nonsingular ($a_0\neq 0$), and write A^{-1} as a polynomial, of degree at most $n-1$, in A. We record this latter fact as a corollary to (2.4.2).

2.4.4 Corollary. If $A\in M_n$ is nonsingular, then there is a polynomial $q(t)$ (whose coefficients depend upon A), of degree at most $n-1$, such that $A^{-1}=q(A)$.

Exercise. If two matrices $A, B\in M_n$ are similar, show that any polynomial evaluated at one is similar to the same polynomial evaluated at the other, and, in particular, any polynomial equation satisfied by one is satisfied by the other. Give some thought to the converse: Satisfaction of the same polynomial equations implies similarity – true or false?

2.4.5 Example. We have shown that every matrix $A\in M_n$ satisfies some polynomial equation of degree n. The characteristic polynomial may be used, for example. It is possible for $A\in M_n$ to satisfy a polynomial equation of degree less than n, however. Let

$$A=\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}\in M_3$$

Then A satisfies $q(A)=0$, where $q(t)=t^2-2t+1$ has degree 2.

Exercise. Show that a diagonalizable matrix satisfies a polynomial equation of degree equal to the number of its distinct eigenvalues, and no less. The (monic) polynomial of minimum degree that a matrix satisfies – its minimum polynomial – will be a subject of further study in the next chapter in connection with the Jordan canonical form. *Hint:* Consider $q(t)=(t-\lambda_1)\cdots(t-\lambda_k)$ with $\lambda_i\neq\lambda_j$.

Another use of Schur's result is to make it clear that every matrix is "almost" diagonalizable in two possible interpretations of the phrase. The first says that arbitrarily close to a given matrix there is a diagonalizable matrix, and the second says that any given matrix is similar to an upper triangular matrix whose off-diagonal entries are arbitrarily small.

2.4.6 Theorem. Let $A=[a_{ij}]\in M_n$. For every $\epsilon>0$, there exists a matrix $A(\epsilon)=[a_{ij}(\epsilon)]\in M_n$ that has n distinct eigenvalues (and is therefore diagonalizable) and is such that

$$\sum_{i,j=1}^{n}|a_{ij}-a_{ij}(\epsilon)|^2<\epsilon$$

Proof: Let $U\in M_n$ be unitary and such that $U^*AU=T$ is upper triangular. Let $E=\operatorname{diag}(e_1,e_2,\ldots,e_n)$ in which $e_1,\ldots,e_n$ are numbers chosen so that

$$|e_i|<\left(\frac{\epsilon}{n}\right)^{1/2}$$

and so that the numbers $t_{11}+e_1, t_{22}+e_2,\ldots,t_{nn}+e_n$ are distinct. (Reflect for a moment to see that this can be done.) Then $T+E$ has n distinct eigenvalues: $t_{11}+e_1,\ldots,t_{nn}+e_n$, and so does $A+UEU^*$, which is similar to $T+E$. Let $A(\epsilon)=A+UEU^*$, so that $A-A(\epsilon)=-UEU^*$ and

$$\sum_{i,j}|a_{ij}-a_{ij}(\epsilon)|^2=\sum_{i=1}^{n}|e_i|^2<n\left(\frac{\epsilon}{n}\right)=\epsilon$$

We have used (2.2.2). Therefore, $A(\epsilon)$ satisfies the assertions of the theorem. □

Exercise. Show that the condition $\sum_{i,j}|a_{ij}-a_{ij}(\epsilon)|^2<\epsilon$ in (2.4.6) could be replaced by $\max_{i,j}|a_{ij}-a_{ij}(\epsilon)|<\epsilon$. *Hint:* Apply the theorem with ϵ^2 in place of ϵ and realize that, if a sum of squares is less than ϵ^2, each of the items must be less than ϵ in absolute value.

2.4.7 Theorem. Let $A\in M_n$. For every $\epsilon>0$ there is a nonsingular matrix $S_\epsilon\in M_n$ such that

$$S_\epsilon^{-1}AS_\epsilon = T_\epsilon = [t_{ij}(\epsilon)]$$

is upper triangular and $|t_{ij}(\epsilon)| < \epsilon$ for $1 \le i < j \le n$.

Proof: First apply Schur's theorem to produce a unitary matrix $U \in M_n$ and an upper triangular matrix $T \in M_n$ such that

$$U^*AU = T$$

Define $D_\alpha = \text{diag}(1, \alpha, \alpha^2, \ldots, \alpha^{n-1})$ for a nonzero scalar α and set $t = \max_{i<j}|t_{ij}|$. Assume that $\epsilon < 1$, since it certainly suffices to prove the statement in this case. If $t \le 1$, let $S_\epsilon = UD_\epsilon$, and, if $t > 1$, let $S_\epsilon = UD_{1/t}D_\epsilon$. In either case, the appropriate S_ϵ substantiates the claim of the theorem. If $t \le 1$, for example, a simple calculation reveals that $t_{ij}(\epsilon) = t_{ij}\epsilon^{-i}\epsilon^{j} = t_{ij}\epsilon^{j-i}$, whose absolute value is no more than ϵ^{j-i}, which is, in turn, no more than ϵ if $i < j$. If $t > 1$, on the other hand, the similarity by $D_{1/t}$ simply preprocesses the matrix, producing one in which all off-diagonal entries are no more than 1 in absolute value. □

Exercise. Prove the following variation upon (2.4.7): If $A \in M_n$ and $\epsilon > 0$, there is a nonsingular $S_\epsilon \in M_n$ such that $S_\epsilon^{-1}AS_\epsilon = T_\epsilon = [t_{ij}(\epsilon)]$ is upper triangular and $\sum_{j>i}|t_{ij}(\epsilon)| < \epsilon$. *Hint:* Apply (2.4.7) with $[2/n(n-1)]\epsilon$ in place of ϵ.

An extension of Schur's theorem, easily proved from it, is an important step toward the Jordan canonical form, to come in the next chapter.

2.4.8 **Theorem.** Suppose that $A \in M_n$ has eigenvalues λ_i with multiplicity n_i, $i = 1, \ldots, k$, and that $\lambda_1, \ldots, \lambda_k$ are distinct. Then A is similar to a matrix of the form

$$\begin{bmatrix} T_1 & & & 0 \\ & T_2 & & \\ & & \ddots & \\ 0 & & & T_k \end{bmatrix}$$

where $T_i \in M_{n_i}$ is upper triangular with all diagonal entries equal to λ_i, $i = 1, \ldots, k$. If $A \in M_n(\mathbf{R})$ and if all the eigenvalues of A are real, then the same result holds, and the similarity matrix may be taken to be real.

Proof: First apply Schur's theorem (2.3.1) to exhibit (unitary) similarity to an upper triangular matrix $T = [t_{rs}]$, and suppose that we have arranged that all the λ_1 terms come first, the λ_2 terms next, and so on, on

the diagonal of T. We next perform a sequence of simple (nonunitary) similarities on T that produce the desired above-diagonal 0's, without changing the diagonal or the upper triangular structure of T. Let E_{rs} be that matrix in M_n with a 1 in the r, s position and 0's elsewhere. Notice that, for $r \ne s$ and α any scalar, $I + \alpha E_{rs}$ is nonsingular and $(I + \alpha E_{rs})^{-1} = I - \alpha E_{rs}$. Furthermore, straightforward calculation reveals that the similarity by $I + \alpha E_{rs}$ for $r < s$,

$$(I + \alpha E_{rs})^{-1}T(I + \alpha E_{rs}) = (I - \alpha E_{rs})T(I + \alpha E_{rs})$$

changes entries of T only in the rth row, to the right of column s, and in the sth column above the rth row,

$$\begin{bmatrix} \uparrow \!\! \rightarrow \\ r, s \end{bmatrix}$$

and it replaces t_{rs} by

$$t_{rs} + \alpha(t_{rr} - t_{ss})$$

Thus, if $t_{rr} \ne t_{ss}$, the r, s entry may be made 0 by choosing

$$\alpha = \frac{-t_{rs}}{(t_{rr} - t_{ss})}$$

without otherwise altering the relevant structure. Now consider the sequence of positions: $(n-1, n)$; $(n-2, n-1)$, $(n-2, n)$; $(n-3, n-2)$, $(n-3, n-1)$, $(n-3, n)$; $(n-4, n-3)\ldots$ in T. Make each of these 0, in turn, via a similarity of the indicated sort, *if* $t_{rr} \ne t_{ss}$, and notice that no previously created 0 entry will be disturbed. The resulting matrix will be similar to A and will have the desired form. □

Exercise. Show that if $A \in M_n(\mathbf{R})$ and all its eigenvalues are real, all operations necessary in the proof of (2.4.8) may be carried out in real arithmetic. Thus, in this event, the block diagonal matrix guaranteed by the theorem, and the similarity necessary to achieve it, may be taken to be real.

Remark: Suppose a given matrix $A \in M_n$ is upper triangular and, by a permutation similarity if necessary, suppose it has been reduced to the form

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1k} \\ & \ddots & \vdots \\ 0 & & A_{kk} \end{bmatrix}$$

in which each diagonal block A_{ii} is upper triangular and has only λ_i on the diagonal, and suppose $\lambda_i \neq \lambda_j$ if $i \neq j$. The algorithm used to prove Theorem (2.4.8) shows that A is similar to

$$\begin{bmatrix} A_{11} & & 0 \\ & \ddots & \\ 0 & & A_{kk} \end{bmatrix}$$

That is, in this situation, all the off-diagonal blocks may be replaced with zero blocks and similarity is preserved. Because unitary similarity preserves the sum of squares of the absolute values of the entries, notice that it is not possible to accomplish this result with a *unitary* similarity if any off-diagonal block A_{ij} is nonzero.

We now use the commuting families version (2.3.3) of Schur's theorem to show that the eigenvalues "add" – in some order – for commuting matrices.

2.4.9 Theorem. Let $A, B \in M_n$ have eigenvalues $\alpha_1, \ldots, \alpha_n$ and $\beta_1, \ldots, \beta_n$, respectively. If A and B commute, there is a permutation $i_1, \ldots, i_n$ of the indices $1, \ldots, n$, such that the eigenvalues of $A+B$ are $\alpha_1+\beta_{i_1}, \alpha_2+\beta_{i_2}, \ldots, \alpha_n+\beta_{i_n}$. In particular, $\sigma(A+B) \subseteq \sigma(A)+\sigma(B)$ if A and B commute.

Proof: If A and B commute, they may be simultaneously upper-triangularized according to (2.3.3); that is, there is a unitary $U \in M_n$ such that

$$U^*AU = T \quad \text{and} \quad U^*BU = R$$

are both upper triangular, with diagonal entries $\alpha_1, \ldots, \alpha_n$ and $\beta_{i_1}, \ldots, \beta_{i_n}$, respectively. Observe that

$$U^*(A+B)U = T+R$$

has diagonal entries and therefore eigenvalues

$$\alpha_1+\beta_{i_1}, \alpha_2+\beta_{i_2}, \ldots, \alpha_n+\beta_{i_n}$$

These must also be the eigenvalues of $A+B$ since $A+B$ is similar to $T+R$. □

2.4.10 Example. Note that, even when A and B commute, not necessarily all numbers of the form $\alpha_i+\beta_j$ occur as eigenvalues of $A+B$. Consider the diagonal matrices

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 3 & 0 \\ 0 & 4 \end{bmatrix}$$

and realize that $1+4=5 \notin \{4,6\} = \sigma(A+B)$. Thus $\sigma(A+B)$ is contained in, but is not generally equal to, $\sigma(A)+\sigma(B)$ when A and B commute.

2.4.11 Example. If A and B do *not* commute, it is difficult to say much about $\sigma(A+B)$ in terms of $\sigma(A)$ and $\sigma(B)$. In particular, $\sigma(A+B)$ need not be contained in $\sigma(A)+\sigma(B)$. Let

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$$

Then $\sigma(A+B) = \{-1, 1\}$, while $\sigma(A) = \sigma(B) = \{0\}$.

2.4.12 Example. Is the converse of (2.4.9) true? If the eigenvalues of A and B add, in some order, need A and B commute? The answer is no, even if the eigenvalues of αA and βB add, in some order, for all scalars α and β. This is an interesting phenomenon, and the characterization of such pairs of matrices is an unsolved problem! Let

$$A = \begin{bmatrix} 1 & 4 & 5 \\ 0 & 2 & 6 \\ 0 & 0 & 3 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 2 & 1 & 2 \\ 0 & 3 & 3 \\ 0 & 0 & 4 \end{bmatrix}$$

The eigenvalues add, but A and B do not commute. Clearly, simultaneous upper triangularizability is sufficient for the additivity of eigenvalues, but it, too, is not necessary. And, of course, upper triangular matrices need not commute.

2.4.13 Corollary. Suppose that $A, B \in M_n$ are commuting matrices with eigenvalues $\alpha_1, \ldots, \alpha_n$ and $\beta_1, \ldots, \beta_n$, respectively. If $\alpha_i \neq -\beta_j$, $i, j = 1, \ldots, n$, then $A+B$ is nonsingular.

Exercise. Verify (2.4.13) using (2.4.9).

Exercise. Show that for any pair $A, B \in M_n$ (commuting or noncommuting) the sum of the eigenvalues of $A+B$ is the sum of the eigenvalues of A plus the sum of the eigenvalues of B. *Hint:* What is $\operatorname{tr}(A+B)$?

We have considered simultaneous diagonalization of diagonalizable matrices, for which commutativity is an easily verified necessary and sufficient condition. We have also considered simultaneous triangularization, for which commutativity is a sufficient but not necessary condition. Since it is sometimes useful to be able to show that two given matrices cannot be simultaneously triangularized, we are interested in stronger

necessary conditions than additivity of the eigenvalues. The following example points the way toward such conditions.

2.4.14 **Example.** Let

$$A=\begin{bmatrix}0&1&0\\0&0&-1\\0&0&0\end{bmatrix}\quad\text{and}\quad B=\begin{bmatrix}0&0&0\\1&0&0\\0&1&0\end{bmatrix}$$

Both A and B have the eigenvalue 0 with multiplicity 3, as does any linear combination $aA+bB$, so the eigenvalues add and there is, on these grounds, reason to believe that A and B might be simultaneously triangularizable. But if there were some nonsingular $S\in M_3$ such that SAS^{-1} and SBS^{-1} were both upper triangular, then the eigenvalues of $(SAS^{-1})(SBS^{-1})=SABS^{-1}$ would have to be products, in some order, of the eigenvalues of A and B. But the set of eigenvalues of AB is $\{-1,0,1\}$, which is not contained in the set product of $\{0\}$ and $\{0\}$. We conclude that A and B are not simultaneously upper triangularizable.

Exercise. Verify the assertions in the preceding example, including the fact that if $C,D\in M_n$ are both upper triangular, then the eigenvalues of CD are products of eigenvalues of C and D in some order; that is, $\sigma(CD)\subseteq\sigma(C)\sigma(D)$.

Simultaneous upper triangularizability by a not necessarily unitary similarity [but see Section (2.6)] is completely characterized by the following theorem of McCoy, whose proof we omit. Recall that we may speak of a polynomial in any number of variables; it is simply a linear combination of products of powers of the variables. If the variables are noncommuting, different powers of the same variables may occur in a given product with products of powers of other variables in between.

2.4.15 **Theorem.** Let $A,B\in M_n$, with $\sigma(A)=\{\alpha_1,\dots,\alpha_n\}$ and $\sigma(B)=\{\beta_1,\dots,\beta_n\}$, including multiplicities. There is a nonsingular $S\in M_n$ such that both $S^{-1}AS$ and $S^{-1}BS$ are upper triangular if and only if there is a permutation $i_1,\dots,i_n$ of the indices $1,2,\dots,n$ such that $\sigma(p(A,B))=\{p(\alpha_j,\beta_{i_j}):j=1,\dots,n\}$ for all polynomials $p(t,s)$ with complex coefficients in two (noncommuting) variables.

Exercise. Verify that the polynomial condition in (2.4.15) is necessary if A and B are simultaneously triangularizable. Show that, if $A,B\in M_n$ commute, then $\sigma(p(A,B))=\{p(\alpha_j,\beta_{i_j}):j=1,\dots,n\}$ for all polynomials p in two variables. How is Example (2.4.14) covered by the theorem?

Remark: The result of Theorem (2.4.15) is valid for matrices and polynomials over an arbitrary field as long as the field contains the eigenvalues of the matrices; it holds for simultaneous triangularization of $k=3,4,\dots$ matrices (the condition then involves polynomials in k variables); and it even holds for a restricted subset of the eigenvalues, that is, $p(\alpha_j,\beta_{i_j})\in\sigma(p(A,B))$, $j=1,\dots,r$, for all polynomials $p(s,t)$ if and only if A and B are simultaneously similar to block triangular matrices with $\alpha_1,\dots,\alpha_r$ and $\beta_{i_1},\dots,\beta_{i_r}$ comprising 1-by-1 blocks in corresponding positions somewhere on the diagonals of the block triangular matrices that are simultaneously similar to A and B, respectively.

Problems

1. Suppose $A,B\in M_n$ commute and have eigenvalues $\alpha_1,\dots,\alpha_n$ and $\beta_1,\dots,\beta_n$, respectively. (a) Show that the eigenvalues of AB are $\alpha_1\beta_{i_1}$, $\alpha_2\beta_{i_2},\dots,\alpha_n\beta_{i_n}$ for some permutation $i_1,\dots,i_n$ of the indices $1,\dots,n$. (b) If $p(t,s)$ is a polynomial in two variables, show also that $p(A,B)$ has eigenvalues $p(\alpha_1,\beta_{i_1}),\dots,p(\alpha_n,\beta_{i_n})$. (c) Finally, show that the weaker assumption of simultaneous upper triangularizability is sufficient for the previous conclusions; commutativity is not necessary.

2. If $A\in M_n$, show that the rank of A is not less than the number of nonzero eigenvalues of A. *Hint:* Show that the rank of an upper triangular matrix is at least as great as the number of nonzero main diagonal entries, and then use Schur's theorem (2.3.1). Use the example

$$A=\begin{bmatrix}0&1\\0&0\end{bmatrix}$$

to explain why the rank of A could be greater than the number of nonzero eigenvalues.

3. The purpose of this problem is to show that the Cayley–Hamilton theorem holds for matrices whose entries come from any *commutative ring,* not just from the complex field. A commutative ring is a mathematical structure in which all the axioms of a field are satisfied *except* for the existence of multiplicative inverses. Thus, there are commutative operations of "addition" and "multiplication" that obey the usual properties of associativity and distributivity. We also explicitly assume that there is a multiplicative unit in the ring; that is, there is an element "1" such that $1a=a$ for all a in the ring. One example of a ring that is not necessarily a field is Z_k = the integers modulo k. In Z_k, "addition" and "multiplication" are done as usual, but the result is taken modulo k; Z_k is a field if and only if k is a prime. Another example is the set of polynomials in k formal indeterminants with complex coefficients.

(a) Recall that if $A \in M_n$, then $\operatorname{adj} A \in M_n$ is the unique matrix whose i, j entry is the j, i cofactor of A (0.8.2). Show that the fundamental identity

$$A(\operatorname{adj} A) = (\operatorname{adj} A)A = (\det A)I$$

is just an expression of Laplace's expansion of the determinant of A by cofactors and the fact that $\det A = 0$ if any two rows or columns of A are equal. Observe that this formula involves multiplications and additions only, not division. Argue that it is valid for matrices whose entries come from any commutative ring.

(b) Use (a) to show that

$$(tI - A)[\operatorname{adj}(tI - A)] = [\operatorname{adj}(tI - A)](tI - A) = \det(tI - A)I = p_A(t)I$$

for any $A \in M_n$, or even for any n-by-n matrix A whose entries come from a commutative ring. Show that the matrix $\operatorname{adj}(tI - A)$ is a matrix whose entries are polynomials in t of degree at most $n-1$, and hence it can be written as

$$\operatorname{adj}(tI - A) = A_{n-1}t^{n-1} + A_{n-2}t^{n-2} + \cdots + A_1 t + A_0$$

where the coefficients A_k are n-by-n matrices whose entries are polynomial functions of the entries of A. The polynomial $p_A(t)$ is the characteristic polynomial of A.

(c) Show that

$$\begin{aligned} t^k I - A^k &= (tI - A)(It^{k-1} + At^{k-2} + \cdots + A^{k-2}t + A^{k-1}) \\ &\equiv (tI - A)G_k(A, t) \end{aligned}$$

for $k = 0, 1, 2, \ldots$ if A is an n-by-n matrix whose entries come from a commutative ring. Conclude that

$$t^k I = It^k = A^k + (tI - A)G_k(A, t), \qquad k = 0, 1, 2, \ldots$$

(d) Let $p_A(t) = a_n t^n + a_{n-1}t^{n-1} + \cdots + a_1 t + a_0 = \det(tI - A)$ be the characteristic polynomial of A (with $a_n = 1$) and observe that it is well defined for any n-by-n matrix A whose entries come from a commutative ring. Use (c) to show that

$$\begin{aligned} p_A(t)I &= \sum_{k=0}^{n} a_k t^k I = \sum_{k=0}^{n} a_k [A^k + (tI - A)G_k(A, t)] \\ &= p_A(A) + (tI - A)G(A, t) \end{aligned}$$

where

$$G(A, t) = \sum_{k=0}^{n} a_k G_k(A, t)$$

is a polynomial of degree at most $n-1$ in t with matrix coefficients whose entries are polynomial functions of the entries of A. Now use (b) to show that

$$\begin{aligned} p_A(A) &= p_A(t)I - (tI - A)G(A, t) \\ &= (tI - A)\operatorname{adj}(tI - A) - (tI - A)G(A, t) \\ &= (tI - A)H(A, t) \equiv Q_A(t) \end{aligned}$$

where $H(A, t) = B_{n-1}t^{n-1} + B_{n-2}t^{n-2} + \cdots + B_1 t + B_0$ and each B_k is an n-by-n matrix whose entries are polynomial functions of the entries of A that do not depend on t. Thus, $Q_A(t)$ is a polynomial in t of degree at most n with matrix coefficients.

(e) Now evaluate $Q_A(A)$ and conclude that $p_A(A) = 0$.

4. Let $A \in M_n$ be a nonsingular matrix. Show that any matrix that commutes with A also commutes with A^{-1}. *Hint:* Use (2.4.4); give a direct argument as well.

5. Use (2.3.1) to show that if $A \in M_n$ has eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, then

$$\sum_{i=1}^{n} \lambda_i^k = \operatorname{tr} A^k, \qquad k = 1, 2, \ldots$$

6. Show that for

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} -2 & 1 & 2 \\ -1 & -2 & -1 \\ 1 & 1 & 1 \end{bmatrix}$$

$\sigma(aA + bB) = \{a - 2b, 2a - 2b, 3a + b\}$ for all scalars $a, b \in \mathbf{C}$, but A and B are not simultaneously similar to upper triangular matrices. $\sigma(AB) = ?$

7. Use the criterion in Problem 6 of Section (2.3) to show that the two matrices in Example (2.4.14) cannot be simultaneously upper triangularized. Apply the same test to the two matrices in Problem 6.

8. An observation in the spirit of McCoy's theorem (2.4.15) can sometimes be useful in showing that two matrices are not unitarily equivalent. Let $p(t, s)$ be a polynomial with complex coefficients in two noncommuting variables, and let $A, B \in M_n$ be unitarily equivalent with $A = UBU^*$ for some unitary $U \in M_n$. Show that $p(A, A^*) = Up(B, B^*)U^*$. Conclude that if A and B are unitarily equivalent, then $\operatorname{tr} p(A, A^*) = \operatorname{tr} p(B, B^*)$ for every complex polynomial $p(t, s)$ in two noncommuting variables. How is this related to Theorem (2.2.6)?

9. Let $A \in M_n$, $B \in M_m$ be given and suppose A and B have no eigenvalues in common; that is, $\sigma(A) \cap \sigma(B)$ is empty. Use the Cayley–Hamilton

theorem (2.4.2) to show that the equation $AX-XB=0$, $X\in M_{n,m}$, has only the solution $X=0$. Deduce from this fact that the equation $AX-XB=C$ has a unique solution $X\in M_{n,m}$ for each given $C\in M_{n,m}$. *Hint:* If $AX=XB$, show inductively that $A^kX=XB^k$ for all $k=1,2,\dots$ and hence $p(A)X=Xp(B)$ for any polynomial $p(t)$. Choose $p(t)$ to be the characteristic polynomial of A to obtain $p_A(A)X=0=Xp_A(B)$. Since $p_A(B)=(B-\lambda_1 I)\cdots(B-\lambda_n I)$, where $\lambda_1,\dots,\lambda_n$ are the eigenvalues of A, the matrix $p_A(B)$ is nonsingular and $Xp_A(B)=0$ has only the solution $X=0$. Existence of a solution to $AX-XB=C$ for any given right-hand side follows from uniqueness of the solution to the homogeneous equation and (0.5*k* and *l*) applied to the linear transformation $X\to T(X)=AX-XB$ on $M_{n,m}$.

10. Use Problem 9 to give a proof of Theorem (2.4.8) that requires at most $k-1$ reduction steps. *Hint:* Write A as

$$A=\begin{bmatrix} A_{11} & A_{12} & \dots & A_{1k} \\ 0 & A_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & A_{kk} \end{bmatrix}=\begin{bmatrix} A_{11} & R_1 \\ 0 & T \end{bmatrix}$$

in which each A_{ii} is upper triangular with only λ_i on the main diagonal and $R_1=[A_{12}\dots A_{1k}]$. Consider

$$S=\begin{bmatrix} I & X \\ 0 & I \end{bmatrix},\qquad S^{-1}=\begin{bmatrix} I & -X \\ 0 & I \end{bmatrix}$$

where X is the same size as R_1. Show that

$$S^{-1}AS=\begin{bmatrix} A_{11} & 0 \\ 0 & T \end{bmatrix}$$

provided X is chosen so that $A_{11}X-XT=-R_1$. Continue with the successive rows and conclude that A is similar to $\operatorname{diag}(A_{11},A_{22},\dots,A_{kk})$.

11. Let $A,B\in M_n$ be given and consider the *commutator* $C=AB-BA$. Show that $\operatorname{tr} C=0$. Consider $A=\left[\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right]$ and $B=\left[\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right]$ and show that a commutator need not be nilpotent; that is, some eigenvalues of a commutator can be nonzero, even though the sum of the eigenvalues must be zero.

12. Let $A,B\in M_n$, let $C=AB-BA$, and suppose that A commutes with C. Show that C must be nilpotent. Comment on the example in Problem 11. *Hint:* Why is there a nonsingular $S\in M_n$ such that $SCS^{-1}=\operatorname{diag}(C_{11},C_{22},\dots,C_{kk})\equiv C_1$, where each $C_{ii}\in M_{n_i}$ is upper triangular, $n_1+n_2+\cdots+n_k=n$, $\sigma(C_{ii})=\{\lambda_i\}$ for $i=1,2,\dots,k$, and $\lambda_i\neq\lambda_j$ if $i\neq j$? Let $A_1\equiv SAS^{-1}$, $B_1\equiv SBS^{-1}$, and write $A_1=(A_{ij})$ and $B_1=(B_{ij})$ in block form conformal with the block diagonal form of C_1. Show that $A_1C_1=C_1A_1$ and use Problem 9 to show that $A_{ij}=0$ if $k>1$ and $i\neq j$. Then each $C_{ii}=A_{ii}B_{ii}-B_{ii}A_{ii}$ has $\operatorname{tr} C_{ii}=0$ and hence $\lambda_i=0$ and $k=1$.

13. Using the notation of Problem 9, use (2.4.9) to give another proof of the fact that the equation $AX-XB=C$ has a unique solution for every $C\in M_n$ if A and B have no eigenvalues in common. *Hint:* Consider the linear transformations $T_1,T_2:M_{n,m}\to M_{n,m}$ defined by $T_1(X)=AX$, $T_2(X)=XB$. Show that T_1 and T_2 commute, and deduce from (2.4.9) that the eigenvalues of T are differences of eigenvalues of T_1 and T_2. Argue that λ is an eigenvalue of T_1 if and only if there is a nonzero $X\in M_{n,m}$ such that $AX-\lambda X=0$, which can happen if and only if λ is an eigenvalue of A [consider the nonzero column(s) of X]. The sets of eigenvalues of T_1 and A are therefore the same, and similarly for T_2 and B. Thus, T is nonsingular if A and B have no eigenvalues in common. If x is an eigenvector of A corresponding to the eigenvalue λ and y is an eigenvector of B^T corresponding to the eigenvalue μ, consider $X=xy^T$, show that $T(X)=(\lambda-\mu)X$, and conclude that the set of eigenvalues of T consists of all possible differences of eigenvalues of A and B.

14. Let $\mathfrak{F}=\{A_i:i\in\mathfrak{I}\}\subset M_n$ be a commuting family. Show that $\mathfrak{F}$ can be simultaneously upper-triangularized in such a way that any one given member is reduced to the special form in (2.4.8) and the other members are reduced to conformal block diagonal upper triangular form. That is, for each given $A_0\in\mathfrak{F}$, show that there is a nonsingular $S\in M_n$ such that $A_i=S\operatorname{diag}(T_1^{(i)},\dots,T_k^{(i)})S^{-1}$ for all $i\in\mathfrak{I}$, each $T_j^{(i)}\in M_{n_j}$ is upper triangular for $j=1,2,\dots,k$, $n_1+n_2+\cdots+n_k=n$, and all $i\in\mathfrak{I}$, and all the main diagonal entries of each $T_j^{(0)}$ are λ_j, with $\lambda_j\neq\lambda_i$ if $j\neq i$. *Hint:* Choose S so that $S^{-1}A_0S$ has the special block diagonal upper triangular form in (2.4.8). Note that the family $\{S^{-1}A_iS:i\in\mathfrak{I}\}$ is commutative. Partition each $S^{-1}A_iS$ conformally to the block form of $S^{-1}A_0S$ and employ commutativity and the result of Problem 9 or 13 (as in Problem 12) to show that all the off-diagonal blocks of each $S^{-1}A_iS$ must vanish. Now use (2.3.3) on the k families of correspondingly placed diagonal blocks. Except for $S^{-1}A_0S$, there is, of course, no guarantee that the eigenvalues of one diagonal block of $S^{-1}A_iS$ are all equal or that different diagonal blocks have disjoint spectra.

Further Readings and Notes. Theorem (2.4.15) and its generalizations were proved by N. H. McCoy, "On the Characteristic Roots of Matrix Polynomials," *Bull. Amer. Math. Soc.* 42 (1936), 592–600. See also

T. S. Motzkin and O. Taussky, "Pairs of Matrices with Property L," *Trans. Amer. Math. Soc.* 73 (1952), 108–114, where the relation between eigenvalues and linear combinations is discussed. A pair $A, B \in M_n$ such that $\sigma(aA+bB) = \{a\alpha_j + b\beta_{i_j} : j = 1, \ldots, n\}$ for all $a, b \in \mathbf{C}$ is said to have *property L*, and the condition of (2.4.15) is called *property P*. Clearly property P implies property L and not conversely. The weaker property L is not fully understood, although it is known, for example, that a pair of normal matrices [see (2.5)] with property L must commute and must therefore be simultaneously unitarily diagonalizable.

2.5 Normal matrices

The class of normal matrices, which arises naturally in the context of unitary equivalence, is important throughout matrix analysis and generalizes unitary, real symmetric, and Hermitian matrices.

2.5.1 **Definition.** A matrix $A \in M_n$ is said to be *normal* if $A^*A = AA^*$, that is, if A commutes with its Hermitian adjoint.

Exercise. Show that $A \in M_n$ is normal if and only if every matrix that is unitarily equivalent to A is normal. The class of normal matrices is closed under unitary equivalence.

2.5.2 **Examples.**

(a) Since $U^*U = I = UU^*$ if U is unitary, all unitary matrices are normal.
(b) Since $A^*A = AA^*$ trivially if $A^* = A$, all Hermitian matrices are normal.
(c) If $A \in M_n$ is such that $A^* = -A$, A is called *skew-Hermitian*. In this event, $A^*A = -A^2 = AA^*$, so that all skew-Hermitian matrices are also normal.
(d) $A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ is normal, but it does not fall into any of the above categories.

Exercise. Characterize the normal matrices in $M_2(\mathbf{R})$ in terms of relations among the entries. Present the answer in terms of the categories (2.5.2a, b, and c). *Hint:* If $A \in M_2(\mathbf{R})$ is normal, show that either $A = A^T$ or $A = -A^T$ if at least one entry of A is zero. If all entries of A are nonzero, show that either $A = A^T$ or $AA^T = aI$ for some $a > 0$.

Exercise. Give an example of a 2-by-2 real matrix that is *not* normal. Give an example of a real 2-by-2 matrix that is normal but is not symmetric, skew-symmetric, or orthogonal.

Exercise. Show that each of the categories (2.5.2a, b, and c) is itself closed under unitary equivalence.

Exercise. Show that a diagonal Hermitian matrix must have real entries and a diagonal skew-Hermitian matrix must have pure imaginary entries.

2.5.3 **Definition.** If $A \in M_n$ is unitarily equivalent to a diagonal matrix, A is said to be *unitarily diagonalizable*, with a similar definition for *orthogonally diagonalizable.* Note that "unitarily (or orthogonally) diagonalizable" implies diagonalizable (but not conversely).

Exercise. Review the proof of (1.3.7) and conclude that $A \in M_n$ is unitarily diagonalizable if and only if there is a set of n orthonormal vectors in $\mathbf{C}^n$, each of which is an eigenvector of A.

We next catalog the most fundamental facts about normal matrices. The equivalence of (a) and (b) in the following theorem is often called the *spectral theorem for normal matrices.*

2.5.4 **Theorem.** If $A = [a_{ij}] \in M_n$ has eigenvalues $\lambda_1, \ldots, \lambda_n$, the following statements are equivalent:

(a) A is normal;
(b) A is unitarily diagonalizable;
(c) $\sum_{i,j=1}^{n} |a_{ij}|^2 = \sum_{i=1}^{n} |\lambda_i|^2$; and
(d) There is an orthonormal set of n eigenvectors of A.

Proof: We suppose throughout that $T = [t_{ij}] \in M_n$ is an upper triangular matrix which is unitarily equivalent to A, as guaranteed by Schur's theorem (2.3.1); that is, $T = U^*AU$ for some unitary $U \in M_n$. Since T is unitarily equivalent to A, the statement (a) is equivalent to the normality of T. We proceed by showing that (a) is equivalent to (b), (b) is equivalent to (c), and (b) is equivalent to (d).

To show that (a) implies (b), we use a calculation. If A is normal, then T is normal. But a triangular normal matrix must be diagonal, as may be seen by equating the diagonal entries of T^*T and TT^*. The fact that the 1, 1 entry of T^*T is the same as that of TT^* means that

$$\bar{t}_{11}t_{11} = t_{11}\bar{t}_{11} + \sum_{j=2}^{n} t_{1j}\bar{t}_{1j} = |t_{11}|^2 + \sum_{j=2}^{n} |t_{1j}|^2$$

This means that $0 = \sum_{j=2}^{n} |t_{1j}|^2$, a sum of nonnegative terms, each of which must then be 0. We conclude that

$$t_{1j} = 0, \ j = 2, \ldots, n$$

The fact that the 2, 2 entries of T^*T and TT^* are the same then means that

$$\bar{t}_{22}t_{22} = t_{22}\bar{t}_{22} + \sum_{j=3}^{n} t_{2j}\bar{t}_{2j} = |t_{22}|^2 + \sum_{j=3}^{n} |t_{2j}|^2$$

and we conclude for the same reason as above that

$$t_{2j} = 0, \qquad j = 3, \ldots, n$$

In the same manner, assuming we have verified that

$$t_{ij} = 0, \qquad j > i \quad \text{and} \quad i = 1, \ldots, k-1$$

we may conclude that

$$t_{ij} = 0, \qquad j > i \quad \text{for} \quad i = k$$

Arguing upon each successive diagonal entry in turn, we conclude, finally, that

$$t_{ij} = 0, \qquad j > i \quad \text{for} \quad i = 1, \ldots, n$$

and, since

$$t_{ij} = 0, \qquad j < i \quad \text{for} \quad i = 1, \ldots, n$$

because T is upper triangular, we have that T is diagonal and (b) holds. Since diagonal matrices are clearly normal and unitary equivalence preserves normality, (b) implies (a) also.

For the equivalence of (b) and (c), we appeal to (2.2.2). Since the diagonal entries of any diagonalization of A are the eigenvalues $\lambda_1, \ldots, \lambda_n$ (in some order), (2.2.2) allows us to deduce (c) from (b). On the other hand, since the λ_i, $i = 1, \ldots, n$, are the diagonal entries of T (in some order), (2.2.2) also means that

$$\sum_{i,j=1}^{n} |a_{ij}|^2 = \sum_{i=1}^{n} |\lambda_i|^2 + \sum_{i<j} |t_{ij}|^2$$

But (c) means that

$$\sum_{i<j}^{n} |t_{ij}|^2 = 0$$

or that T is diagonal, from which (b) follows.

The equivalence of (b) and (d) is the content of the exercise preceding this theorem. □

Exercise. If $T \in M_n$ is triangular, and if the ith diagonal entry of T^*T is the same as that of TT^*, $i = 1, \ldots, n$, show that T is diagonal. Explain why this fact, together with the fact that normality is invariant under unitary similarity, is the basic reason why a normal matrix is unitarily diagonalizable.

Exercise. Show that a normal matrix is nondefective (the geometric multiplicity is the same as the algebraic multiplicity for each eigenvalue).

Exercise. Show that if $A \in M_n$ is normal, then $x \in \mathbf{C}^n$ is a right eigenvector of A corresponding to the eigenvalue λ of A if and only if x is a left eigenvector of A corresponding to λ; that is, $Ax = \lambda x$ is equivalent to $x^*A = \lambda x^*$. *Hint:* Normalize x and write $A = U\Lambda U^*$ with x as the first column of U. Then what is A^*? A^*x? See Problem 20 at the end of this section for another proof.

Exercise. If $A \in M_n$ is normal, and if x and y are eigenvectors corresponding to distinct eigenvalues, show that x and y are orthogonal. *Hint:* If $Ax = \lambda x$, $Ay = \mu y$, show that $\mu x^*y = x^*(Ay) = (A^*x)^*y = (\bar{\lambda}x)^*y = \lambda x^*y$. If $\lambda \neq \mu$, conclude that $x^*y = 0$. See Problem 21 for another proof.

If the eigenvalues of a normal matrix are known, it can be unitarily diagonalized via the following conceptual prescription. Determine each eigenspace and find an orthonormal basis for it (using the Gram–Schmidt procedure, for example). Since the dimension of each eigenspace is equal to the multiplicity of the corresponding eigenvalue, and since A is normal, the union of these bases will be an orthonormal basis for the entire space. Arraying these vectors as the columns of a unitary matrix produces the desired diagonalizing transformation.

We next note that commuting normal matrices may be simultaneously unitarily diagonalized.

2.5.5 Theorem. If $\mathfrak{N} \subseteq M_n$ is a commuting family of normal matrices, then $\mathfrak{N}$ is simultaneously unitarily diagonalizable; that is, there is a single unitary similarity that transforms each matrix in $\mathfrak{N}$ into a diagonal matrix.

Exercise. Use (2.3.3) and the fact that a triangular normal matrix must be diagonal to prove (2.5.5). Explain how both the hypothesis and conclusion of (2.5.5) are stronger than those of (1.3.19).

The application of (2.5.4) to the special case of Hermitian matrices yields a fundamental result, often called the *spectral theorem for Hermitian matrices.*

2.5.6 **Theorem.** If $A \in M_n$ is Hermitian, then

(a) All eigenvalues of A are real; and
(b) A is unitarily diagonalizable.

If $A \in M_n(\mathbf{R})$ is symmetric, then A is real orthogonally diagonalizable.

Proof: A diagonal Hermitian matrix must have real diagonal entries, so (a) follows from (b) and the fact that the set of Hermitian matrices is closed under unitary equivalence. Statement (b) follows from (2.5.4) because Hermitian matrices are normal. If $A \in M_n(\mathbf{R})$ is symmetric, then it is Hermitian, but notice that all calculations necessary to diagonalize A take place over the real field. Since the eigenvalues of A are real, the corresponding eigenvectors may be taken to be real. □

It is important to realize, in contrast to the discussion of diagonalizability in Chapter 1, that distinctness of eigenvalues or the like is no longer important in (2.5.4) and (2.5.6), and diagonalizability need not be assumed in (2.5.5). A full linearly independent set of eigenvectors (in fact an orthonormal set) is structurally guaranteed. This is one reason why Hermitian and normal matrices are so important and have such pleasant properties.

We conclude with analogs of (2.5.4) and (2.5.5) for real normal matrices. Such matrices are normal and therefore can be diagonalized by a not necessarily real unitary similarity, but what is the best form such a matrix can be put into by a *real orthogonal* similarity? Since a real normal matrix might have no real eigenvalues whatever, there is no guarantee it can be diagonalized with a real similarity. On the other hand, any real matrix can be put into a special block upper triangular form by a real orthogonal similarity (2.3.4), and this suggests what to do if the matrix is also normal. Our proof uses (2.3.4) in the same way that (2.3.1) was used in the proof of (2.5.4). The following lemma disposes of a slight technical complication that was not present in the proof of (2.5.4).

2.5.7 **Lemma.** If $A \in M_n$ is Hermitian and $x^*Ax \geq 0$ for all $x \in \mathbf{C}^n$, then all the eigenvalues of A are nonnegative. If, in addition, $\operatorname{tr} A = 0$, then $A = 0$.

Proof: Use (2.5.6) to write $A = U\Lambda U^*$, where $U = [u_1\, u_2 \cdots u_n] \in M_n$ is unitary and $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Then $\Lambda = U^*AU$, so $\lambda_k = u_k^* A u_k \geq 0$

by hypothesis, and hence all $\lambda_k \geq 0$. Finally, $\operatorname{tr} A = \operatorname{tr} U\Lambda U^* = \operatorname{tr} \Lambda U^* U = \operatorname{tr} \Lambda = \lambda_1 + \cdots + \lambda_n$, so if $\operatorname{tr} A = 0$ and all $\lambda_k \geq 0$, we must have all $\lambda_k = 0$, $\Lambda = 0$, and $A = U\Lambda U^* = U0U^* = 0$. □

2.5.8 **Theorem.** Let $A \in M_n(\mathbf{R})$. Then A is normal if and only if there is a real orthogonal matrix $Q \in M_n(\mathbf{R})$ such that

$$Q^TAQ = \begin{bmatrix} A_1 & & & 0 \\ & A_2 & & \\ & & \ddots & \\ 0 & & & A_k \end{bmatrix} \in M_n(\mathbf{R}), \qquad 1 \leq k \leq n \tag{2.5.9}$$

where each A_j is either a real 1-by-1 matrix or is a real 2-by-2 matrix of the form

$$A_j = \begin{bmatrix} \alpha_j & \beta_j \\ -\beta_j & \alpha_j \end{bmatrix} \tag{2.5.10}$$

Proof: A direct calculation shows that every matrix of the form (2.5.10) is normal $[A_jA_j^T = \operatorname{diag}(\alpha_j^2 + \beta_j^2, \alpha_j^2 + \beta_j^2) = A_j^TA_j]$, so any direct sum of the form (2.5.9) must also be normal. For the forward implication, we invoke (2.3.4) to show that it suffices to prove the theorem for a real normal matrix of the form (2.3.5). Since the main diagonal blocks in (2.3.5) may be arranged in any prescribed order, we may assume that

$$A = \begin{bmatrix} R & A_{01} & A_{02} & \cdots & A_{0k} \\ & A_{11} & A_{12} & \cdots & A_{1k} \\ & & A_{22} & \cdots & A_{2k} \\ & 0 & & \ddots & \vdots \\ & & & & A_{kk} \end{bmatrix} \in M_n(\mathbf{R}) \tag{2.5.11}$$

is normal with

$$R = \begin{bmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix} \in M_p(\mathbf{R})$$

upper triangular, $A_{01}, A_{02}, \ldots, A_{0k} \in M_{p,2}(\mathbf{R})$, and $A_{ij} \in M_2(\mathbf{R})$ for $i, j = 1, 2, \ldots, k$ and $j \geq i$. We shall show that R is diagonal and that $A_{ij} = 0$ for all $j > i$.

Equating the first p-by-p main diagonal block entries in the identity $A^TA = AA^T$ corresponding to the block R in (2.5.11) gives the identity

$$R^TR = RR^T + A_{01}A_{01}^T + \cdots + A_{0k}A_{0k}^T \tag{2.5.12}$$

Observe that every matrix $B\in M_p(\mathbf{C})$ of the form $B=EE^*$ for some $E\in M_{p,q}$ is Hermitian and has the property that $x^*Bx=x^*EE^*x=(E^*x)^*(E^*x)\geq 0$ for all $x\in\mathbf{C}^n$, and a sum of such matrices has the same property. Since

$$\operatorname{tr} R^TR=\operatorname{tr} RR^T$$

on general principles, and

$$\operatorname{tr} R^TR=\operatorname{tr} RR^T+\operatorname{tr} A_{01}A_{01}^T+\cdots+\operatorname{tr} A_{0k}A_{0k}^T$$

from (2.5.12), we see that

$$0=\operatorname{tr} A_{01}A_{01}^T+\cdots+\operatorname{tr} A_{0k}A_{0k}^T$$

By Lemma (2.5.7) and the above observation applied to the real matrix $B=A_{0j}A_{0j}^*=A_{0j}A_{0j}^T$, we have $\operatorname{tr} A_{0j}A_{0j}^T\geq 0$. Since the sum is zero, each term is zero, and hence $A_{0j}A_{0j}^T=0$ for $j=1,\ldots,k$. The ith main diagonal entry of $A_{0j}A_{0j}^T$ is the sum of the squares of the (real) elements in the ith row of A_{0j}, so all these elements must be zero, all $A_{0j}=0$, $j=1,2,\ldots,k$, and (2.5.12) reduces to

$$R^TR=RR^T$$

But we have already shown in the proof of (2.5.4) that a triangular normal matrix must be diagonal, so we have $R=\operatorname{diag}(\lambda_1,\ldots,\lambda_p)$, as asserted.

Equating the main diagonal 2-by-2 block entries in the identity $A^TA=AA^T$ corresponding to the block A_{11} in (2.5.11) and using the fact that all $A_{0j}=0$ for $j=1,2,\ldots,k$ gives the identity

$$A_{11}^TA_{11}=A_{11}A_{11}^T+A_{12}A_{12}^T+\cdots+A_{1k}A_{1k}^T \tag{2.5.13}$$

But $\operatorname{tr}(A_{11}^TA_{11})=\operatorname{tr}(A_{11}A_{11}^T)$, so

$$\operatorname{tr}(A_{12}A_{12}^T)+\cdots+\operatorname{tr}(A_{1k}A_{1k}^T)=0,\qquad \operatorname{tr}(A_{1j}A_{1j}^T)\geq 0$$

$$\operatorname{tr}(A_{1j}A_{1j}^T)=0,\qquad A_{1j}A_{1j}^T=0,\qquad \text{and}\qquad A_{1j}=0$$

for $j=2,3,\ldots,k$, using Lemma (2.5.7) in the same way as before. Thus, (2.5.13) reduces to $A_{11}^TA_{11}=A_{11}A_{11}^T$; that is, the 2-by-2 block A_{11} is normal.

Examining each successive main diagonal 2-by-2 block entry of the identity $A^TA=AA^T$ corresponding to the block A_{ii} in (2.5.11) for $i=2,3,\ldots,k-1$, and employing this same argument, we conclude that all the off-diagonal blocks are zero as asserted, and all the diagonal blocks A_{ii} are normal.

We have shown that a real orthogonal similarity that reduces a real normal matrix to the form (2.3.5) actually reduces it to the block diagonal

form (2.5.9). We complete the proof by showing that all the diagonal blocks have the form (2.5.10).

If $A_{jj}=\begin{bmatrix} a & b\\ c & d\end{bmatrix}\in M_2(\mathbf{R})$ is normal, then equating the 1, 1 and 1, 2 entries of the identity $A_{jj}^TA_{jj}=A_{jj}A_{jj}^T$ gives the identities

$$b^2=c^2,\qquad \text{so}\quad c=\pm b$$

and

$$ac+bd=ab+cd,\qquad \text{so}\quad 2b(a-d)=0\quad \text{if}\quad c=-b$$

The cases $c=+b$ and $b=0$ can be excluded since A_{jj} would then be real symmetric and would have only real eigenvalues; by our construction, the blocks A_{jj} have conjugate pairs of nonreal eigenvalues. Thus, $c=-b$, $a=d$, and A_{jj} must have the form (2.5.10). A calculation shows that a real matrix $\begin{bmatrix} a & b\\ -b & a\end{bmatrix}$ has a pair of conjugate complex eigenvalues $\lambda=a+ib$ and $\bar\lambda=a-ib$. □

As a consequence of this theorem for real normal matrices, we can easily deduce the real canonical forms for real matrices that are symmetric, skew-symmetric, or orthogonal.

2.5.14 Corollary. Let $A\in M_n(\mathbf{R})$. Then

(a) $A=A^T$ if and only if there is a real orthogonal matrix $Q\in M_n(\mathbf{R})$ such that

$$Q^TAQ=\begin{bmatrix} \lambda_1 & & 0\\ & \ddots & \\ 0 & & \lambda_n\end{bmatrix}\qquad \text{with all}\quad \lambda_i\in\mathbf{R}$$

(b) $A=-A^T$ if and only if there is a real orthogonal matrix $Q\in M_n(\mathbf{R})$ such that

$$Q^TAQ=\begin{bmatrix} 0 & & & & & 0\\ & \ddots & & & & \\ & & 0 & & & \\ & & & A_1 & & \\ & & & & \ddots & \\ 0 & & & & & A_k\end{bmatrix}$$

where each $A_j\in M_2(\mathbf{R})$ has the form

$$A_j=\begin{bmatrix} 0 & \beta_j\\ -\beta_j & 0\end{bmatrix}$$

and

(c) $AA^T = I$ if and only if there is a real orthogonal matrix $Q \in M_n(\mathbf{R})$ such that

$$Q^TAQ = \begin{bmatrix} \lambda_1 & & & & & 0 \\ & \ddots & & & & \\ & & \lambda_p & & & \\ & & & A_1 & & \\ & & & & \ddots & \\ 0 & & & & & A_k \end{bmatrix}$$

where each $\lambda_j = \pm 1$ and each $A_j \in M_2(\mathbf{R})$ has the form

$$A_j = \begin{bmatrix} \cos\theta_j & \sin\theta_j \\ -\sin\theta_j & \cos\theta_j \end{bmatrix}, \qquad \theta_j \in \mathbf{R}$$

Proof: In each case, the hypothesis guarantees that A is a real normal matrix, so A can be written in the form (2.5.9) and (2.5.10). If $A = A^T$, then each $A_j = A_j^T$, so all $\beta_j = 0$ and Q^TAQ is diagonal. If $A = -A^T$, then each $\lambda_j = -\lambda_j$ and each $A_j = -A_j^T$, so all $\lambda_j = 0$ and all $\alpha_j = 0$. If $AA^T = I$, then each $\lambda_j\lambda_j = 1$ and each $A_jA_j^T = I$, so all $\lambda_j^2 = 1$ and all $\alpha_j^2 + \beta_j^2 = 1$; we have $\lambda_j = \pm 1$ and $\alpha_j = \cos\theta_j$, $\beta_j = \sin\theta_j$ in this case. □

If one has a family of commuting real normal matrices, they might not be simultaneously real diagonalizable, but they can all be brought simultaneously into block diagonal form (2.5.9).

2.5.15 **Theorem.** If $\mathfrak{N} \subseteq M_n(\mathbf{R})$ is a commuting family of real normal matrices, then there is a single real orthogonal matrix Q such that Q^TAQ is of the form (2.5.9) and (2.5.10) for all $A \in \mathfrak{N}$.

Proof: Use (2.3.6) to reduce every member of $\mathfrak{N}$ to the form (2.3.5) via one real orthogonal similarity Q. The argument in the proof of (2.5.8) shows that they then have the form (2.5.9). □

Problems

It is possible to accumulate a much longer list than (2.5.4) of conditions on $A \in M_n$ that are equivalent to normality. Several more are included among the problems.

1. Show that $A \in M_n$ is normal if and only if the Euclidean length of Ax is the same as that of A^*x for all $x \in \mathbf{C}^n$. Recall that $(y^*y)^{1/2}$ is the Euclidean length of $y \in \mathbf{C}^n$.

2. Show that a normal matrix is unitary if and only if all its eigenvalues have absolute value 1.

3. Show that a normal matrix is Hermitian if and only if all its eigenvalues are real.

4. Show that a normal matrix is skew-Hermitian if and only if all its eigenvalues are pure imaginary (have real part equal to 0).

5. If $A \in M_n$ is skew-Hermitian (respectively Hermitian), show that iA is Hermitian (respectively skew-Hermitian).

6. Show that $A \in M_n$ is normal if and only if it commutes with some normal matrix with distinct eigenvalues.

7. Consider matrices $A \in M_n$ of the form $A = B^{-1}B^*$ for a nonsingular $B \in M_n$, as in Theorem (2.1.9). (a) Show that A is unitary if and only if B is normal. (b) If B has the form $B = HNH$, where N is normal and H is Hermitian (and both are nonsingular), show that A is similar to a unitary matrix.

8. Define $H(A) = \frac{1}{2}(A + A^*)$, the *Hermitian part,* and $S(A) = \frac{1}{2}(A - A^*)$, the *skew-Hermitian* part, of $A \in M_n$. Then $A = H(A) + S(A)$. Show that A is normal if and only if $H(A)$ and $S(A)$ commute.

9. If two normal matrices commute, show that their product is normal, but show by example that the product of two normal matrices can be normal even if the two factors do not commute.

10. In the notation of Problem 8, show that A is normal if every eigenvector of $H(A)$ is an eigenvector of $S(A)$ (respectively A).

11. For any complex number $z \in \mathbf{C}$, show that there is some $\theta \in \mathbf{R}$ such that $\bar{z} = e^{i\theta}z$. Notice that $[e^{i\theta}] \in M_1$ is a unitary matrix. What do diagonal unitary matrices $U \in M_n$ look like?

12. Generalize Problem 11 to show that if $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \in M_n$, then there is a diagonal unitary matrix U such that $\bar{\Lambda} = U\Lambda = \Lambda U$.

13. Use Problem 12 to show that a matrix $A \in M_n$ is normal if and only if there is a unitary matrix $V \in M_n$ such that $A^* = AV$. How is this related to Problem 7?

14. If $A \in M_n(\mathbf{R})$ and if all the eigenvalues of A are real, show that A is normal if and only if it is symmetric.

15. Show that two normal matrices of the same size are similar (in fact, are unitarily equivalent) if and only if they have the same characteristic

polynomial. Is this true for matrices that are not normal? *Hint:* Consider $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$.

16. If $A, B \in M_n$ are normal, show that AB need not be normal and that the nonsingular normal matrices of a given size do not form a multiplicative group. The unitary normal matrices do form a group, however. Do the nonsingular Hermitian matrices form a multiplicative group?

17. If $A \in M_n$ is normal and $p(t)$ is a given polynomial, use Definition (2.5.1) to show that $p(A)$ is normal. Give another proof of this fact using (2.5.4).

18. If $A \in M_n$ has the property that $p(A)$ is normal for some nonzero polynomial $p(t)$, is A normal? *Hint:* Consider $A = \begin{bmatrix} 0 & 1 \\ 2 & 0 \end{bmatrix}$ and A^2.

19. Let $A \in M_n$ and $a \in \mathbf{C}$ be given. Show that A is normal if and only if $A + aI$ is normal.

20. Let $A \in M_n$ be a normal matrix and suppose $x \in \mathbf{C}^n$ is a vector such that $Ax = \lambda x$. Use Problems 1 and 19 to show that $A^*x = \bar{\lambda}x$. *Hint:* If the Euclidean length of the vector $(A - \lambda I)x$ is zero, argue that the Euclidean length of $(A - \lambda I)^*x$ is also zero.

21. Use (2.5.4) to show that if $A \in M_n$ is normal and if $Ax = \lambda x$ and $Ay = \mu y$ with $\lambda \neq \mu$, then x and y are orthogonal. *Hint:* Write $A = U\Lambda U^*$ with $U \in M_n$ unitary and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$. Let $U^*x = x' = [x_i']$ and $U^*y = y' = [y_i']$. Show that $\Lambda x' = \lambda x'$ and deduce from this that $x_i' = 0$ for every index i such that $\lambda_i \neq \lambda$, and similarly for y'. Show that x' and y' are orthogonal and conclude that x and y are orthogonal.

22. Use (2.5.6) to show that the characteristic polynomial of a Hermitian matrix A must have real coefficients even if not all of the entries of A are real.

23. Show that $\begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}$ and $\begin{bmatrix} i & i \\ i & 1 \end{bmatrix}$ are both complex symmetric matrices $(A = A^T)$, but one is normal and the other is not. There is, therefore, a crucial difference between real symmetric matrices and complex symmetric matrices [see Section (4.4)].

24. If $A \in M_n$ is both normal and nilpotent, show that $A = 0$.

25. Let $A \in M_n$ be a given matrix. Show that A is normal if and only if there is a polynomial $p(t)$ of degree at most $n-1$ such that $A^* = p(A)$. *Hint:* If $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$, use Lagrange interpolation to construct a polynomial $p(t)$ such that $p(\Lambda) = \bar{\Lambda}$ and then invoke (2.5.4). How does this "explain" why a normal matrix commutes with its adjoint? If, in addition, A is real, show that the Lagrange interpolation polynomial $p(\cdot)$ such that $A^* = p(A)$ has real coefficients and $A^T = p(A)$. Thus, a real normal matrix A has $A^T = p(A)$ for a real polynomial $p(\cdot)$. See (0.9.11.4).

26. Give an example of a real normal matrix that is unitarily similar to a diagonal matrix but is not real orthogonally similar to a diagonal matrix. Show that a real matrix A is real orthogonally similar to a diagonal matrix if and only if A is symmetric $(A = A^T)$.

27. Show that a given matrix $A \in M_n$ is normal if and only if

$$(Ax)^*(Ay) = (A^*x)^*(A^*y)$$

for all $x, y \in \mathbf{C}^n$. Geometrically, this means that the angle between Ax and Ay is the same as the angle between A^*x and A^*y for all $x, y \in \mathbf{C}^n$. How is this related to Problem 1?

28. If $A \in M_n$ is normal, show that $Ax = 0$ if and only if $A^*x = 0$. This means that the null space of A is the same as the null space of A^*. Consider $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$ to show that this is not true in general.

29. Consider the system of linear equations $Ax = y$, where $A \in M_n$ and $y \in \mathbf{C}^n$ are given, and suppose A is singular. The given system has a (nonunique) solution if and only if $y^*z = 0$ for every $z \in \mathbf{C}^n$ such that $A^*z = 0$ [see (0.6.6)]. If A is normal, however, show that the given system has a solution if and only if $y^*w = 0$ for every $w \in \mathbf{C}^n$ such that $Aw = 0$; that is, y is orthogonal to the null space of A. If one wants to find *all* solutions to a singular system $Ax = y$, explain why it is computationally more economial to do so if A is normal than if it is not.

30. Let $n_1, n_2, \ldots, n_k$ be given positive integers and let $A_j \in M_{n_j}$, $j = 1, 2, \ldots, k$. Show that the direct sum $A = A_1 \oplus \cdots \oplus A_k$ is normal if and only if each A_j is normal.

31. Show that two normal matrices are similar if and only if they are unitarily equivalent. *Hint:* Show that $U\Lambda U^*$ and $V\Lambda V^*$ are unitarily equivalent if U and V are unitary. Give an example of two (nonnormal) matrices that are similar but not unitarily equivalent.

32. If $A \in M_3(\mathbf{R})$ is a real orthogonal matrix, observe that A has either one or three real eigenvalues. If it has a positive determinant, use (2.5.14) to show that it is orthogonally equivalent to the direct sum of $[1] \in M_1$ and a plane rotation. Discuss the geometrical interpretation of this as a rotation by an angle θ around some fixed axis passing through the origin in $\mathbf{R}^3$. This is part of Euler's theorem in mechanics: Every motion of a rigid body is the composition of a translation and a rotation about some axis.

33. If $\mathfrak{F} \subseteq M_n$ is a commuting family of normal matrices, show that there exists a single Hermitian matrix B such that for each $A_\alpha \in \mathfrak{F}$ there is a polynomial $p_\alpha(t)$ of degree at most $n-1$ such that $A_\alpha = p_\alpha(B)$. Notice that B is fixed for all of $\mathfrak{F}$ but the polynomial may depend on the element of $\mathfrak{F}$. *Hint:* Let $U \in M_n$ be a unitary matrix that simultaneously diagonalizes every member of $\mathfrak{F}$, let $B = U \operatorname{diag}(1, 2, \ldots, n) U^*$, let $A_\alpha = U \Lambda_\alpha U^*$ with $\Lambda_\alpha = \operatorname{diag}(\lambda_1^{(\alpha)}, \ldots, \lambda_n^{(\alpha)})$, and take $p_\alpha(t)$ to be the Lagrange interpolation polynomial such that $p_\alpha(k) = \lambda_k^{(\alpha)}$, $k = 1, 2, \ldots, n$.

34. Show that $A \in M_n$ is normal if and only if every eigenvector of A is also an eigenvector of A^*. *Hint:* Let $U \in M_n$ be a unitary matrix whose first column is an eigenvector of A (and, therefore, of A^*). Inspect both U^*AU and $U^*A^*U = (U^*AU)^*$ and continue.

35. Verify the following improvement of (2.2.8) in case $A, B \in M_n$ are normal: A is unitarily equivalent to B if and only if $\operatorname{tr} A^k = \operatorname{tr} B^k$, $k = 1, 2, \ldots, n$. *Hint:* Use Problem 15, and Problem 12 in Section (1.2).

36. Let $A \in M_n(\mathbf{R})$ and suppose $AA^T = A^TA$, so A is a real normal matrix. If the eigenvalues of AA^T are all distinct, show that A must be symmetric. *Hint:* Use Theorem (2.5.8).

2.6 *QR* factorization and algorithm

A particular means for calculating a specific unitary Schur upper triangularization (2.3.1) of a given matrix $A \in M_n$, and a popular numerical method for calculating eigenvalues (under some assumptions) is called the *QR* algorithm. It is based on the so-called *QR* factorization of a general matrix $A \in M_{n,m}$.

2.6.1 **Theorem** (*QR* factorization). If $A \in M_{n,m}$ and $n \geq m$, there is a matrix $Q \in M_{n,m}$ with orthonormal columns and an upper triangular matrix $R \in M_m$ such that $A = QR$. If $m = n$, Q is unitary; if in addition A is nonsingular, then R may be chosen so that all its diagonal entries are positive, and in this event, the factors Q and R are both unique. If $A \in M_{n,m}(\mathbf{R})$, then both Q and R may be taken to be real.

Proof: If $A \in M_{n,m}$ and rank $A = m$, the *QR* factorization of A is just a description, in matrix notation, of the result of applying the Gram–Schmidt process (0.6.4) to the columns of A, which comprise an independent set in $\mathbf{C}^n$. A natural extension of the Gram–Schmidt algorithm permits the same description to apply to the general case in which the columns of A may be dependent. Let $A = [a_1 \ldots a_m]$ be written in partitioned form in terms of its columns $a_i \in \mathbf{C}^n$. If $a_1 = 0$, set $q_1 = 0$; otherwise set $q_1 = a_1/(a_1^* a_1)^{1/2}$. For each $k = 2, 3, \ldots, m$, compute

$$y_k = a_k - \sum_{i=1}^{k-1} (q_i^* a_k) q_i$$

just as in the ordinary Gram–Schmidt process. If $y_k = 0$ (which happens if and only if a_k is a linear combination of $a_1, a_2, \ldots, a_{k-1}$), set $q_k = 0$; otherwise set $q_k = y_k/(y_k^* y_k)^{1/2}$. The vectors $q_1, \ldots, q_m$ are then an orthogonal set, each element of which is either a unit vector or the zero vector. Each vector q_j is a linear combination of $a_1, \ldots, a_j$, and the construction ensures that, conversely, each column a_j is a linear combination of $q_1, \ldots, q_j$. Thus, scalars r_{kj} exist such that

$$a_j = \sum_{k=1}^{j} r_{kj} q_k, \qquad j = 1, 2, \ldots, m \tag{2.6.2}$$

If we set $r_{kj} = 0$ for all $k > j$ and set $r_{ij} = 0$ for all $j = 1, 2, \ldots, m$ for each i such that $q_i = 0$, the upper triangular matrix $R = [r_{ij}] \in M_m$ and the vectors $q_1, q_2, \ldots, q_m$ are determined, via the outlined procedure, by $a_1, a_2, \ldots, a_m$. The matrix $Q \equiv [q_1 \ldots q_m] \in M_{n,m}$ has orthogonal columns (some of which may be zero), and (2.6.2) says that $A = QR$.

If rank $A = m$, Q has orthonormal columns, and hence a factorization of the desired form has been achieved. In particular, if $m = n$ and A is nonsingular, then Q must be unitary by (2.1.4e) and the diagonal entries of the nonsingular matrix $R = Q^*A$ must be nonzero. In this event, because R is required to be upper triangular, q_1 is a scalar multiple of a_1, and, for $i = 2, 3, \ldots, m$, q_i lies in the one-dimensional space that is the orthogonal complement of the span of $a_1, \ldots, a_{i-1}$ with respect to the span of $a_1, \ldots, a_i$. Therefore, each q_i is uniquely determined up to a factor of scale of absolute value 1. Thus, replacement of R by $R' \equiv \operatorname{diag}(|r_{11}|/r_{11}, \ldots, |r_{mm}|/r_{mm})R$ and replacement of Q by $Q' \equiv Q \operatorname{diag}(r_{11}/|r_{11}|, \ldots, r_{mm}/|r_{mm}|)$ gives the unique factorization $A = Q'R'$ promised in the statement of the theorem.

If the columns of A are not independent, take the (orthonormal) set of nonzero columns of Q and extend it to an orthonormal basis of $\mathbf{C}^n$; denote the new vectors obtained in this way by $z_1, z_2, \ldots, z_p$. Now replace the first zero column of Q by z_1, replace the second zero column by z_2, and so on until all zero columns have been replaced in this way; denote the resulting matrix by Q'. Then Q' has orthonormal columns and $QR = Q'R$ because the new columns of Q' are matched by zero rows of R. Then $A = Q'R$ is a factorization of the desired form.

If A is real, notice that all the necessary operations may be carried out in real arithmetic, so the factors obtained are real. □

Exercise. If $A \in M_{n,m}$ and $n \le m$, show that A may be factored as $A = LP$, in which $L \in M_n$ is lower triangular and $P \in M_{n,m}$ has orthonormal rows, and that there are statements for this factorization parallel to the remaining ones in (2.6.1).

Exercise. Show that any matrix $B \in M_n$ of the form $B = A^*A$, $A \in M_n$, may be written as $B = LL^*$, where $L \in M_n$ is lower triangular with nonnegative diagonal entries. Show that this factorization is unique if A is nonsingular. This is called the *Cholesky factorization* of B; every positive definite matrix may be factored in this way (see Chapter 7). *Hint:* Write $A = QR$.

The QR factorization has considerable numerical significance [e.g., (2.6.3)], but it is also of interest as a theoretical tool. For example, the upper triangularizability of a given matrix $A \in M_n$ by unitary similarity follows immediately from its upper triangularizability by ordinary similarity. If $S^{-1}AS = T$ is upper triangular and $S = QR$ as in (2.6.1), then $R^{-1}Q^*AQR = T$ and $Q^*AQ = RTR^{-1}$, which, as a product of upper triangular matrices, is again upper triangular. It follows in the same way that simultaneous triangularization theorems such as (2.4.15) are actually simultaneous unitary triangularization theorems. That is, if a given family of matrices in M_n is simultaneously triangularizable by similarity, it is also simultaneously triangularizable by unitary equivalence.

We next state the QR algorithm for eigenvalue calculation and briefly indicate some of its features without proof.

2.6.3 **QR algorithm.** Let $A_0 \in M_n$ be given. Write $A_0 = Q_0R_0$, where Q_0 and R_0 are as guaranteed in (2.6.1), and define $A_1 = R_0Q_0$. Again, write $A_1 = Q_1R_1$, with Q_1 unitary and R_1 upper triangular, and continue. In general, factor $A_k = Q_kR_k$ and define $A_{k+1} = R_kQ_k$.

Exercise. Show that each A_k produced by the QR algorithm is unitarily equivalent to A_0, $k = 1, 2, \ldots$.

Under certain circumstances (for example, if all the eigenvalues of A_0 have distinct absolute values), the QR iterates A_k will converge to an upper triangular matrix as $k \to \infty$. Since this upper triangular matrix is unitarily equivalent to A_0, the eigenvalues of A_0 are revealed.

If A_0 is real, then the QR iterates A_k may be calculated using real arithmetic. If A_0 has any nonreal eigenvalues, there is no hope that the QR iterates will converge to an upper triangular matrix, since this upper triangular limit must be real. Under certain circumstances, however, the iterates A_k may be chosen so that they converge to a real block upper triangular matrix with 1-by-1 and 2-by-2 main diagonal blocks. A sufficient condition for this to occur is that all the eigenvalues of A_0 have distinct moduli, except for the two eigenvalues in each non-real complex conjugate pair, which have the same modulus. Since the eigenvalues of a block triangular matrix are the union of the sets of eigenvalues of the diagonal blocks, the eigenvalues of A_0 are revealed as the 1-by-1 diagonal entries of the block triangular limit of the QR iterates A_k, together with the (complex conjugate pairs of) eigenvalues of the 2-by-2 diagonal blocks of the limit, which may be calculated easily using real arithmetic and the quadratic formula.

2.6.4 **Example.** That the QR algorithm need not always converge to a triangular matrix is indicated by the following example. Let

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

Then, $\sigma(A) = \{\pm i\}$ and the eigenvalues do not have distinct moduli. If $A_0 = A$, a possible sequence for the algorithm is

$$A_0 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$A_1 = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

$$A_2 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = A_0$$

Another is

$$A_0 = A_0\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = A_0$$

In either case, cycling occurs and the sequence $\{A_k\}$ does not converge to an upper triangular matrix. There is a choice of $\{A_k\}$ that converges to a block upper triangular matrix, however.

Problems

1. Let $x_1, \ldots, x_m$ be given independent vectors in $\mathbf{C}^n$, and let $X \equiv [x_1\, x_2 \ldots x_m] \in M_{n,m}$. Suppose that the Gram–Schmidt process as described in (0.6.4) is performed on the vectors $x_1, \ldots, x_m$ to produce an orthonormal set $z_1, \ldots, z_m$, and let $Z \equiv [z_1 \ldots z_m] \in M_{n,m}$. (a) For $k = 1, 2, \ldots, m$, let $Z_k \equiv [z_1\, z_2 \ldots z_k\, x_{k+1}\, x_{k+2} \ldots x_m]$, where z_k is the unit vector produced

by the kth step of the Gram–Schmidt process, so that $Z_m = Z$. Show that $Z_1 = X\Delta_1$, $Z_2 = Z_1\Delta_2, \ldots, Z_m = Z_{m-1}\Delta_m$, where each Δ_i is a nonsingular upper triangular matrix of the form $\Delta_i = I +$an upper triangular matrix in which all columns but the ith are zero. (b) Let $T_k \equiv \Delta_1\Delta_2 \ldots \Delta_k$, $k = 1, 2, \ldots, m$. Observe that T_k is upper triangular and $Z_k = XT_k$, $k = 1, 2, \ldots, m$. Let $T \equiv T_m$, so that $Z = XT$. (c) What is the relationship between this matrix T and the upper triangular matrix R in the proof of Theorem (2.6.1)? (d) Show that the first k columns of Z_j and T_j do not change for $j = k+1, k+2, \ldots, m$, so the kth step of the Gram–Schmidt process produces the kth columns of the final matrices Z and T.

2. Let $x_1, \ldots, x_m$ be given independent vectors in $\mathbf{C}^n$ and let $X \equiv [x_1 \ldots x_m] \in M_{n,m}$. Consider the following algorithm:

I. Set $Z \equiv X$, and denote $Z = [z_1 \ldots z_m]$, so that initially each column $z_i = x_i$, $i = 1, \ldots, m$.

II. For $k = 1, 2, \ldots, m$, do the following:
 (i) First replace the column z_k by $z_k/\langle z_k, z_k\rangle^{1/2}$; then
 (ii) For $j = k+1, k+2, \ldots, m$ replace each column z_j by
 $z_j - \langle z_j, z_k\rangle z_k$.

We use $\langle x, y\rangle \equiv y^*x$ to denote the usual inner product on $\mathbf{C}^n$. (a) Show that the final result of this process is a matrix Z with orthonormal columns and that it is the same matrix Z produced by the Gram–Schmidt process in Problem 1. (b) If Z_k denotes the contents of the matrix Z at the end of the kth step of the algorithm, $k = 1, 2, \ldots, m$, show that $Z_1 = X\Delta_1, Z_2 = Z_1\Delta_2, \ldots, Z_m = Z_{m-1}\Delta_m$, where each Δ_i is a nonsingular upper triangular matrix of the form $\Delta_i = I +$an upper triangular matrix in which all rows but the ith are zero. (c) Let $T_k \equiv \Delta_1\Delta_2 \ldots \Delta_k$, $k = 1, 2, \ldots, m$. Observe that T_k is upper triangular, and $Z_k = XT_k$, $k = 1, \ldots, m$. Show that the first k columns of each T_k are the same as the first k columns of the matrix T_k in Problem 1, though the respective matrices Δ_k and Z_k may be different. Let $T \equiv T_m$. (d) Show that the first k columns of Z_j and T_j do not change for $j = k+1, k+2, \ldots, m$, so the kth step of the algorithm produces the kth column of the final matrices Z and T. This algorithm is known as the *modified Gram–Schmidt* process; it produces the same result as the ordinary Gram–Schmidt process through a rearrangement of the calculation. Although the modified and ordinary Gram–Schmidt algorithms are mathematically equivalent, the modified Gram–Schmidt is preferred for numerical computations because it requires less storage and, in difficult problems in which some columns of X are nearly parallel, it tends to produce a computed Z whose columns are more nearly orthogonal than the Z produced by the ordinary Gram–Schmidt algorithm. Its performance can easily be improved further in difficult problems by a column-pivoting strategy: Before performing step II(i), first choose as z_k a remaining column z_j, $j \geq k$, whose squared length $z_j^*z_j$ is greatest. In numerical computations one actually obtains Δ_i^{-1} at each step (no inversion is required) and accumulates the products of these factors to compute the triangular factor in the QR decomposition of X.

3. Show how the QR factorization may be achieved using a sequence of multiplications by Householder transformations. Show that $n-1$ Householder transformations are necessary and that Q is a product of these. This method is known to be computationally superior to the Gram–Schmidt argument used in the proof of (2.6.1).

4. If the QR algorithm applied to $A_0 \in M_m$ converges to an upper triangular matrix, how may eigenvectors of A_0 be calculated? *Hint:* One needs to solve a (singular) triangular system with zero right-hand side.

5. Let the QR algorithm be applied to a given matrix $A \in M_n$, and let $\{A_k\}$ be the sequence of QR iterates. If the sequence converges, and if $\lim_{k\to\infty} A_k = B$, use the selection principle (2.1.8) to explain carefully why B is unitarily equivalent to A. Why is this important?

Further Readings. For additional references and a detailed description of efficient computational implementations of the Gram–Schmidt, modified Gram–Schmidt, and several other orthogonalization procedures, see pp. 146–169 of [GVl]. For further discussion of the QR algorithm, proofs, and additional references, see the survey by D. Watkins, "Understanding the QR Algorithm," *SIAM Rev.* 24 (1982), 427–440 as well as [Ste].

CHAPTER 3

Canonical forms

3.0 Introduction

When are two given matrices similar? We know that similar matrices have the same trace, determinant, characteristic polynomial, and eigenvalues, but the example

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \tag{3.0.1}$$

shows that two matrices can have all four of these quantities the same without being similar. If there were some nonsingular $S \in M_2$ such that $A = SBS^{-1} = SOS^{-1} = 0$, then we would have a contradiction, since $A \neq 0$.

Exercise. Compute the trace, determinant, characteristic polynomial, and eigenvalues of the two matrices in (3.0.1). Show that $A^2 = 0$.

Since two matrices that look very different can still be similar, one approach to determining whether two given matrices are similar is to have some set of "simple" matrices of prescribed form and then see if both given matrices can be reduced by similarity to one of these "simple" forms. If they can, then they must be similar (because the similarity relation is transitive and symmetric). What "simple" forms would be suitable for this purpose?

Every complex matrix A is (unitarily) similar to an upper triangular matrix whose diagonal entries (the eigenvalues of A) may be arranged in any given order (2.3.1), so two matrices are similar if they are similar to the same upper triangular matrix. However, two upper triangular

matrices with the same main diagonal entries and different off-diagonal entries can still be similar. Thus, if we have succeeded in reducing the two given matrices to two unequal upper triangular matrices with the same main diagonal, we cannot conclude that the matrices are not similar. There is too much freedom here; the $n(n+1)/2$ nonzero entries (or, more precisely, "not necessarily zero" entries) in an upper triangular matrix are too numerous to distinguish similarity. There is no uniqueness about the triangular form.

If the class of upper triangular matrices is too large for our purposes, what about the class of diagonal matrices? If each of two given matrices is similar to a diagonal matrix, then they are indeed similar to each other if and only if the two diagonal matrices have the same main diagonal entries, counting multiplicities but ignoring order. The reason is that one can use a permutation similarity PDP^T to present the main diagonal entries of a diagonal matrix D in any prescribed order. Although this solves the problem of uniqueness that we had with the upper triangular matrices, we now have an existence problem: Not every complex matrix is similar to a diagonal matrix.

Exercise. Show that the matrix A in (3.0.1) is not diagonalizable. *Hint:* If $A=S\Lambda S^{-1}$, then $\Lambda=B$.

If we search for an upper triangular form that is as nearly diagonal as possible but is still attainable by similarity for every matrix, the result is the Jordan canonical form, which we discuss in the next section.

We have been considering similarity of two given matrices $A, B\in M_n$, but there are other equivalence relations of interest in matrix theory. For example, we might be interested in whether A can be transformed into B by a *unitary* similarity, or by applying only elementary row and column operations. If A and B are real, we might want to know if A is similar to B via a real similarity. If A and B are Hermitian, we might want to know if there is a nonsingular $S\in M_n$ such that $A=SBS^*$. If A and B are symmetric, we might want to know if there is a nonsingular $S\in M_n$ such that $A=SBS^T$.

In each of these examples, we have an equivalence relation on a set of matrices and we are interested in whether two given matrices are in the same equivalence class. One approach to solving this problem is to seek a "simple" set of representative matrices of prescribed form, one from each equivalence class, and to try to reduce each given matrix to one of them. If this approach is to be successful at all, each equivalence class must actually contain a representative of the prescribed form (this fails for diagonal matrices under similarity), and it is very desirable to have only one representative (or perhaps a small and readily described set of equivalent representatives) in each class (this fails for upper triangular matrices under similarity). Such a set of representatives is often called a *canonical form,* and we consider several examples in this chapter. Others will arise in particular contexts in later chapters.

3.1 The Jordan canonical form: a proof

The Jordan canonical form is a set of "almost diagonal" matrices, called Jordan matrices, which includes the diagonal matrices. It has the property that every equivalence class (under similarity) of square complex matrices contains a Jordan matrix, and any two Jordan matrices in the same equivalence class are essentially the same in a trivial way. A Jordan matrix that is similar to a given matrix is called *the Jordan canonical form* (or sometimes *the Jordan normal form*) of the matrix. Once one knows the Jordan canonical form of a matrix, all the linear algebraic information about the given matrix (i.e., linear transformation) is known at a glance.

3.1.1 **Definition.** A *Jordan block* $J_k(\lambda)$ is a k-by-k upper triangular matrix of the form

$$J_k(\lambda)=\begin{bmatrix} \lambda & 1 & & & 0 \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ 0 & & & & \lambda \end{bmatrix} \tag{3.1.2}$$

There are $k-1$ terms "+1" in the superdiagonal; the scalar λ appears k times on the main diagonal. All other entries are zero, and $J_1(\lambda)=[\lambda]$. A *Jordan matrix* $J\in M_n$ is a direct sum of Jordan blocks

$$J=\begin{bmatrix} J_{n_1}(\lambda_1) & & & 0 \\ & J_{n_2}(\lambda_2) & & \\ & & \ddots & \\ 0 & & & J_{n_k}(\lambda_k) \end{bmatrix}, \qquad n_1+n_2+\cdots+n_k=n \tag{3.1.3}$$

in which the orders n_i may not be distinct and the values λ_i need not be distinct.

Notice that if each Jordan block $J_{n_i}(\lambda_i)$ in (3.1.3) is one-dimensional, that is, all $n_i=1$ and $k=n$, then the Jordan matrix J is diagonal. If any Jordan block $J_m(\lambda)$ in (3.1.3) has $m>1$, then J is not only not diagonal, it is not even diagonalizable. If $J_m(\lambda)=S\Lambda S^{-1}$ with Λ diagonal, then necessarily $\Lambda=\operatorname{diag}(\lambda,\lambda,\ldots,\lambda)=\lambda I$. Thus, $J_m(\lambda)-\lambda I=S\Lambda S^{-1}-\lambda I=$

$\lambda I - \lambda I = 0$, which is not the case if $m > 1$. There is one eigenvector of J associated with each separate Jordan block; it is the standard basis vector associated with the first diagonal entry of each $J_m(\lambda)$ in J.

The main result of this section is that every complex matrix is similar to an essentially unique Jordan matrix. We shall proceed to this final conclusion in three steps:

Step 1. Observe that every complex matrix is similar to an upper triangular matrix whose eigenvalues appear on the main diagonal in a prescribed order; this is the Schur triangularization theorem (2.3.1).

Step 2. Then show that an upper triangular matrix can be transformed by similarity into a block diagonal matrix in which each individual diagonal block has all its diagonal entries equal [like the Jordan block (3.1.2)]. This is Theorem (2.4.8).

Step 3. Finally, show that an upper triangular matrix whose main diagonal entries are all equal is similar to a direct sum of Jordan blocks (3.1.2).

Once we have proved the last assertion, the reduction of an arbitrary complex matrix to Jordan form follows by combining the similarity transformations required for each step.

We shall also be interested in concluding that if a matrix is real and has only real eigenvalues, then the reduction to Jordan canonical form can be accomplished with a *real* similarity. Toward this end, recall (2.3.1) that if a real matrix A has only real eigenvalues, then there is a real unitary (real orthogonal) matrix U such that U^TAU is upper triangular and hence it has only real entries. Moreover, the proof of (2.4.8) shows that if an upper triangular matrix A is real, then there is a real similarity matrix S such that $S^{-1}AS$ is a (real) block diagonal matrix in which each diagonal block is upper triangular with all its main diagonal entries equal. Thus, it suffices to show that step 3 can be accomplished and that, if we start with a real upper triangular matrix with all main diagonal entries real, then the similarity matrix that reduces it to a direct sum of Jordan blocks may be taken to be real.

The following lemma is helpful in proving that step 3 can be accomplished. Its proof is an entirely straightforward computation.

3.1.4 **Lemma.** Let $k \geq 1$ be given, and consider the Jordan block

$$J_k(0) = \begin{bmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{bmatrix}$$

Then

$$J_k^T(0)J_k(0) = \begin{bmatrix} 0 & 0 \\ 0 & I_{k-1} \end{bmatrix} \quad \text{and} \quad J_k(0)^p = 0 \quad \text{if} \quad p \geq k$$

Moreover,

$$J_k(0)e_{i+1} = e_i \quad \text{for} \quad i = 1, 2, \ldots, k-1 \quad \text{and} \quad [I - J_k^T(0)J_k(0)]x = (x^Te_1)e_1$$

Here, $I_{k-1} \in M_{k-1}$ is the identity matrix, e_i is the ith standard unit basis vector, and $x \in \mathbf{C}^n$.

We now prove that the reduction in step 3 can always be accomplished. Recall that a strictly upper triangular matrix is a triangular matrix with zero entries on the main diagonal. Notice that an upper triangular matrix with equal main diagonal entries is a scalar multiple of the identity plus a strictly upper triangular matrix.

3.1.5 **Theorem.** Let $A \in M_n$ be strictly upper triangular. There is a nonsingular $S \in M_n$ and integers $n_1, n_2, \ldots, n_m$ with $n_1 \geq n_2 \geq \cdots \geq n_m \geq 1$ and $n_1 + n_2 + \cdots + n_m = n$ such that

$$A = S \begin{bmatrix} J_{n_1}(0) & & & 0 \\ & J_{n_2}(0) & & \\ & & \ddots & \\ 0 & & & J_{n_m}(0) \end{bmatrix} S^{-1} \tag{3.1.6}$$

If A is real, the matrix S may be chosen to be real.

Proof: If $n = 1$, $A = [0]$ and the result is trivial. We proceed by induction on n and assume that $n > 1$ and that the result has been proved for all strictly upper triangular matrices of order less than n. Partition A as

$$A = \begin{bmatrix} 0 & a^T \\ 0 & A_1 \end{bmatrix}$$

where $a \in \mathbf{C}^{n-1}$, and $A_1 \in M_{n-1}$ is strictly upper triangular. By the induction hypothesis, there is a nonsingular matrix $S_1 \in M_{n-1}$ such that $S_1^{-1}A_1S_1$ has the desired form (3.1.6); that is,

$$S_1^{-1}A_1S_1 = \begin{bmatrix} J_{k_1} & & & 0 \\ & J_{k_2} & & \\ & & \ddots & \\ 0 & & & J_{k_s} \end{bmatrix} = \begin{bmatrix} J_{k_1} & 0 \\ 0 & J \end{bmatrix} \tag{3.1.7}$$

with $k_1 \geq k_2 \geq \cdots \geq k_s \geq 1$, $k_1 + k_2 + \cdots + k_s = n-1$, $J_{k_i} \equiv J_{k_i}(0)$, and

$$J \equiv \begin{bmatrix} J_{k_2} & & 0 \\ & \ddots & \\ 0 & & J_{k_s} \end{bmatrix} \in M_{n-k_1-1}$$

Notice that no diagonal Jordan block in J has order greater than k_1, so $J^{k_1} = 0$ by Lemma (3.1.4). A simple computation shows that

$$\begin{bmatrix} 1 & 0 \\ 0 & S_1^{-1} \end{bmatrix} A \begin{bmatrix} 1 & 0 \\ 0 & S_1 \end{bmatrix} = \begin{bmatrix} 0 & a^T S_1 \\ 0 & S_1^{-1} A_1 S_1 \end{bmatrix} \tag{3.1.8}$$

Partition $a^T S_1 = [a_1^T a_2^T]$ consistent with the partition of the far right side of (3.1.7); that is, $a_1 \in \mathbf{C}^{k_1}$ and $a_2 \in \mathbf{C}^{n-1-k_1}$, and write (3.1.8) as

$$\begin{bmatrix} 1 & 0 \\ 0 & S_1^{-1} \end{bmatrix} A \begin{bmatrix} 1 & 0 \\ 0 & S_1 \end{bmatrix} = \begin{bmatrix} 0 & a_1^T & a_2^T \\ 0 & J_{k_1} & 0 \\ 0 & 0 & J \end{bmatrix}$$

Now consider the following similarity of this matrix:

$$\begin{bmatrix} 1 & -a_1^T J_{k_1}^T & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} 0 & a_1^T & a_2^T \\ 0 & J_{k_1} & 0 \\ 0 & 0 & J \end{bmatrix} \begin{bmatrix} 1 & a_1^T J_{k_1}^T & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}$$

$$= \begin{bmatrix} 0 & a_1^T(I - J_{k_1}^T J_{k_1}) & a_2^T \\ 0 & J_{k_1} & 0 \\ 0 & 0 & J \end{bmatrix} = \begin{bmatrix} 0 & (a_1^T e_1)e_1^T & a_2^T \\ 0 & J_{k_1} & 0 \\ 0 & 0 & J \end{bmatrix} \tag{3.1.9}$$

where we have used the identity $(I - J_k^T J_k)x = (x^T e_1)e_1$ from Lemma (3.1.4). There are now two possibilities, depending on whether $a_1^T e_1 = 0$ or not.

If $a_1^T e_1 \neq 0$, then

$$\begin{bmatrix} 1/a_1^T e_1 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & (1/a_1^T e_1)I \end{bmatrix} \begin{bmatrix} 0 & (a_1^T e_1)e_1^T & a_2^T \\ 0 & J_{k_1} & 0 \\ 0 & 0 & J \end{bmatrix} \begin{bmatrix} a_1^T e_1 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & a_1^T e_1 I \end{bmatrix}$$

$$= \begin{bmatrix} 0 & e_1^T & a_2^T \\ 0 & J_{k_1} & 0 \\ 0 & 0 & J \end{bmatrix} \equiv \begin{bmatrix} \tilde{J} & e_1 a_2^T \\ 0 & J \end{bmatrix}$$

Notice that

$$\tilde{J} = \begin{bmatrix} 0 & e_1^T \\ 0 & J_{k_1} \end{bmatrix} = J_{k_1+1}(0)$$

is a Jordan block of order k_1+1 with zero main diagonal. Using the property that $\tilde{J}e_{i+1} = e_i$ for $i = 1, 2, \ldots, k_1$, one shows easily that

$$\begin{bmatrix} I & e_2 a_2^T \\ 0 & I \end{bmatrix} \begin{bmatrix} \tilde{J} & e_1 a_2^T \\ 0 & J \end{bmatrix} \begin{bmatrix} I & -e_2 a_2^T \\ 0 & I \end{bmatrix} = \begin{bmatrix} \tilde{J} & -\tilde{J}e_2 a_2^T + e_1 a_2^T + e_2 a_2^T J \\ 0 & J \end{bmatrix}$$

$$= \begin{bmatrix} \tilde{J} & e_2 a_2^T J \\ 0 & J \end{bmatrix}$$

and that one can proceed recursively to compute the series of similarities

$$\begin{bmatrix} I & e_{i+1} a_2^T J^{i-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \tilde{J} & e_i a_2^T J^{i-1} \\ 0 & J \end{bmatrix} \begin{bmatrix} I & -e_{i+1} a_2^T J^{i-1} \\ 0 & I \end{bmatrix} = \begin{bmatrix} \tilde{J} & e_{i+1} a_2^T J^i \\ 0 & J \end{bmatrix},$$

$$i = 2, 3, \ldots$$

Since $J^{k_1} = 0$, we see that after at most k_1 steps in this series of similarities, the off-diagonal term will finally vanish. We conclude that A is similar to the matrix

$$\begin{bmatrix} \tilde{J} & 0 \\ 0 & J \end{bmatrix}$$

which is a strictly upper triangular Jordan matrix of the required form.

If $a_1^T e_1 = 0$, then (3.1.9) shows that A is similar to the matrix

$$\begin{bmatrix} 0 & 0 & a_2^T \\ 0 & J_{k_1} & 0 \\ 0 & 0 & J \end{bmatrix}$$

which is permutation similar to the matrix

$$\begin{bmatrix} J_{k_1} & 0 & 0 \\ 0 & 0 & a_2^T \\ 0 & 0 & J \end{bmatrix} \tag{3.1.10}$$

By the induction hypothesis, there is a nonsingular $S_2 \in M_{n-k_1}$ such that

$$S_2^{-1} \begin{bmatrix} 0 & a_2^T \\ 0 & J \end{bmatrix} S_2 = \hat{J} \in M_{n-k_1}$$

is a Jordan matrix with zero main diagonal. Thus, the matrix (3.1.10), and therefore A itself, is similar to

$$\begin{bmatrix} J_{k_1} & 0 \\ 0 & \hat{J} \end{bmatrix}$$

which is a Jordan matrix of the required form, except that the diagonal Jordan blocks might not be arranged in nonincreasing order. A block permutation similarity, if necessary, will produce the required form.

Finally, observe that if A is real then all the similarities used in this

proof can be chosen to be real, so A is similar via a real similarity to the required Jordan matrix. □

Theorem (3.1.5) essentially completes step 3 of the promised program for exhibiting the Jordan canonical form. Notice that if

$$A = \begin{bmatrix} \lambda & & * \\ & \lambda & \\ 0 & & \ddots \\ & & & \lambda \end{bmatrix}$$

is an upper triangular matrix with all diagonal entries equal to λ, then $A_0 = A - \lambda I$ is strictly upper triangular. If $S \in M_n$ is nonsingular and $S^{-1}A_0S$ is a direct sum of basic Jordan blocks $J_{n_i}(0)$, as guaranteed by (3.1.5), then $S^{-1}AS = S^{-1}A_0S + \lambda I$ is a direct sum of basic Jordan blocks $J_{n_i}(\lambda)$. Steps 1 and 2, carried out in Sections (2.3) and (2.4), together with step 3 explicitly demonstrate the existence half of the *Jordan canonical form theorem:*

3.1.11 **Theorem.** Let $A \in M_n$ be a given complex matrix. There is a nonsingular matrix $S \in M_n$ such that

$$A = S \begin{bmatrix} J_{n_1}(\lambda_1) & & & 0 \\ & J_{n_2}(\lambda_2) & & \\ 0 & & \ddots & \\ & & & J_{n_k}(\lambda_k) \end{bmatrix} S^{-1} = SJS^{-1} \qquad (3.1.12)$$

and $n_1 + n_2 + \cdots + n_k = n$. The Jordan matrix J of A is unique up to permutations of the diagonal Jordan blocks. The eigenvalues λ_i, $i = 1, \ldots, k$ are not necessarily distinct. If A is a real matrix with only real eigenvalues, then the similarity matrix S can be taken to be real.

Proof: We have proved everything except the uniqueness assertion. If $A, B \in M_n$ are similar, then for any scalar $\lambda \in \mathbf{C}$ and any exponent $m = 1, 2, \ldots$, the matrices $(A - \lambda I)^m$ and $(B - \lambda I)^m$ are also similar; in particular, their ranks are equal. Thus, it suffices to show that the set of Jordan blocks (including repetitions) lying on the diagonal of a Jordan matrix $J \in M_n$ is determined completely by the finitely many integers $\operatorname{rank}(J - \lambda I)^m$, $m = 1, 2, \ldots, n$, $\lambda \in \sigma(J)$.

First consider a Jordan block $J_k(\mu) \in M_k$ of the form (3.1.2) for some given $\mu \in \mathbf{C}$ and $m \geq 1$. If $\mu \neq 0$, then $\operatorname{rank} J_k(\mu)^m = \operatorname{rank} J_k(\mu)^{m+1} = k$, so $\operatorname{rank} J_k(\mu)^m - \operatorname{rank} J_k(\mu)^{m+1} = 0$. If $\mu = 0$ and $m \geq k$, then $J_k(0)^m = J_k(0)^{m+1} = 0$, so $\operatorname{rank} J_k(0)^m - \operatorname{rank} J_k(0)^{m+1} = 0$ again. Finally, if $\mu = 0$ and $m < k$, then $\operatorname{rank} J_k(0)^m - \operatorname{rank} J_k(0)^{m+1} = 1$.

Now let $J \in M_n$ be a Jordan matrix of the form (3.1.3), let $\lambda \in \sigma(J)$, and define $r_m(\lambda) \equiv \operatorname{rank}(J - \lambda I)^m$ for $m = 1, \ldots$; set $r_0(\lambda) \equiv n$. It follows from the preceding analysis of the case of one block that the difference $d_m(\lambda) \equiv r_{m-1}(\lambda) - r_m(\lambda)$ is equal to the total number of Jordan blocks $J_k(\lambda)$ in J of all sizes $k \geq m$ and that $d_m(\lambda) = 0$ for all $m > n$. The number of Jordan blocks in J with exact size $k = m$ is therefore equal to $d_m(\lambda) - d_{m+1}(\lambda) = r_{m-1}(\lambda) - 2r_m(\lambda) + r_{m+1}(\lambda)$ for $m = 1, 2, \ldots, n$. □

Exercise. Let $A \in M_n$ have Jordan canonical form J, let λ be an eigenvalue of A with algebraic multiplicity ν, and let b_k denote the number of Jordan blocks $J_k(\lambda)$ of size k in J, $k = 1, \ldots, n$. If $r_m(\lambda) \equiv \operatorname{rank}(A - \lambda I)^m$ for $m \geq 1$ and $r_0(\lambda) \equiv n$, show that: (a) $r_m(\lambda)$ and b_i satisfy the linear equations

$$r_m(\lambda) = n - \nu + \sum_{i=m+1}^{n} (i - m)b_i \qquad m = 0, 1, \ldots, n-1$$

(b) These equations have a unique solution. (c) The solution is $b_m = r_{m-1}(\lambda) - 2r_m(\lambda) + r_{m+1}(\lambda)$, $m = 1, 2, \ldots, n$, where $r_{n+1}(\lambda) = r_n(\lambda) = n - \nu$.

In order to have a standard presentation of the Jordan canonical form (3.1.12), it is conventional to choose some ordering of the distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ of A and to present first all the Jordan blocks corresponding to λ_1, then those corresponding to λ_2, and so on. The Jordan blocks corresponding to each distinct eigenvalue are presented in decreasing (nonincreasing) order with the largest block first, then the next largest, and so on. Since multiple blocks of the same size corresponding to the same eigenvalue are identical, this presentation gives a uniquely determined Jordan canonical form once the ordering of the eigenvalues is given. Every similarity equivalence class of matrices in M_n contains one and only one such Jordan canonical form.

Although our derivation of the Jordan canonical form is an explicit algorithm that can, in principle, be used to compute the Jordan form of a given matrix, it is definitely not recommended for automatic numerical computation by computer. The unfortunate truth is that not only can it produce spurious results, but in fact there exists no numerically stable way to compute Jordan canonical forms. A simple example will make this clear. If $A_\epsilon = \begin{bmatrix} \epsilon & 0 \\ 1 & 0 \end{bmatrix}$ and $\epsilon \neq 0$, then $A_\epsilon = S_\epsilon J_\epsilon S_\epsilon^{-1}$ with $S_\epsilon = \begin{bmatrix} 0 & \epsilon \\ 1 & 1 \end{bmatrix}$ and $J_\epsilon = \begin{bmatrix} 0 & 0 \\ 0 & \epsilon \end{bmatrix}$. If we let $\epsilon \to 0$, then $J_\epsilon \to \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$, which cannot be the Jordan form of the nonzero matrix $A_0 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$. In fact, A_0 has $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ as its Jordan form. Since the Jordan form of a matrix need not be a continuous function of

the entries of the matrix, it is possible that small variations in the entries of a matrix will result in large variations in the entries of the Jordan form. There is no hope of computing such an object in a stable way, so the Jordan canonical form is little used in numerical applications.

Despite this limitation, the Jordan canonical form is well worth knowing and is a rich source of insights. As a matter of general technique, if one has something to prove about matrices it is well to consider first if it can be proved for diagonal matrices and, if this is successful, then to see if some limiting argument may establish the result in general (using the fact that any complex matrix can be approximated arbitrarily closely by a diagonalizable matrix). If this does not work, or if one prefers to avoid an analytical argument, one might next try to prove the result for upper triangular or Jordan matrices.

It is sometimes useful to know that every matrix is similar to a matrix of the form (3.1.12) in which all the "+1" terms in the Jordan blocks are replaced by $\epsilon \neq 0$, and ϵ can be taken to be arbitrarily small.

3.1.13 **Corollary.** Let $A \in M_n$ be a given complex matrix, and let $\epsilon > 0$ be given. Then there exists a nonsingular matrix $S = S(\epsilon) \in M_n$ such that

$$A = S \begin{bmatrix} J_{n_1}(\lambda_1, \epsilon) & & & 0 \\ & J_{n_2}(\lambda_2, \epsilon) & & \\ & & \ddots & \\ 0 & & & J_{n_k}(\lambda_k, \epsilon) \end{bmatrix} S^{-1}, \tag{3.1.14}$$

$$n_1 + n_2 + \cdots + n_k = n$$

and

$$J_m(\lambda, \epsilon) \equiv \begin{bmatrix} \lambda & \epsilon & & 0 \\ & \lambda & \epsilon & \\ & & \ddots & \ddots \\ & & & \epsilon \\ 0 & & & \lambda \end{bmatrix} \in M_m$$

If A is real with real eigenvalues and $\epsilon \in \mathbf{R}$, then S may be taken to be real.

Proof: First find a nonsingular matrix $S_1 \in M_n$ such that $S_1^{-1}AS_1$ is in Jordan canonical form (with a real S_1 if A is real and has real eigenvalues). Then take $D_\epsilon = \text{diag}(1, \epsilon, \epsilon^2, \ldots, \epsilon^{n-1})$ and compute $D_\epsilon^{-1}(S_1^{-1}AS_1)D_\epsilon$. This matrix is of the form (3.1.14), so $S = S(\epsilon) = S_1 D_\epsilon$ meets the requirements of the theorem. □

Problems

1. Supply the computational details to prove Lemma (3.1.4).
2. Carry out the three steps in the proof of (3.1.11) to find the Jordan canonical forms of

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 3 & 1 & 2 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

3. Let $A \in M_n$ be a matrix with complex entries, but with only real eigenvalues. Show that A is similar to a real matrix. Can the similarity matrix be chosen to be real?

Further Readings. Our proof of (3.1.11) is in the spirit of R. Fletcher and D. Sorensen, "An Algorithmic Derivation of the Jordan Canonical Form," *Amer. Math. Monthly* 90 (1983), 12–16, which has additional references. [Ste] discusses the Jordan canonical form from the point of view of numerical computations and gives examples of its sensitivity to perturbations in the entries of the matrix. [Str] presents a nice proof.

3.2 The Jordan canonical form: some observations and applications

3.2.1 **The structure of a Jordan matrix.** The Jordan matrix

$$J = \begin{bmatrix} J_{n_1}(\lambda_1) & & & 0 \\ & J_{n_2}(\lambda_2) & & \\ & & \ddots & \\ 0 & & & J_{n_k}(\lambda_k) \end{bmatrix}, \qquad n_1 + n_2 + \cdots + n_k = n \tag{3.2.1.1}$$

has a definite structure that makes apparent certain basic properties of the matrix and of any matrix that is similar to it.

1. The number k of Jordan blocks (counting multiple occurrences of the same block) is the number of linearly independent eigenvectors of J.
2. The matrix J is diagonalizable if and only if $k = n$.
3. The number of Jordan blocks corresponding to a given eigenvalue is the geometric multiplicity of the eigenvalue, which is the dimension of the associated eigenspace. The sum of the orders of all

the Jordan blocks corresponding to a given eigenvalue is the algebraic multiplicity of the eigenvalue.

4. A Jordan matrix is not completely determined in general by a knowledge of the eigenvalues and their algebraic and geometric multiplicity. One must also know the sizes of the Jordan blocks corresponding to each eigenvalue. The size of the largest Jordan block corresponding to an eigenvalue λ is the multiplicity of λ as a root of the minimal polynomial (3.3.6).
5. The sizes of the Jordan blocks corresponding to a given eigenvalue are determined by a knowledge of the ranks of certain powers. For example, if

$$J = \left[\begin{array}{ccc|cc|cc|c} 2 & 1 & 0 & & & & & \\ 0 & 2 & 1 & & & \multicolumn{3}{c}{0} \\ 0 & 0 & 2 & & & & & \\ \hline & & & 2 & 1 & & & \\ & & & 0 & 2 & & & \\ \hline & & & & & 2 & 1 & \\ & 0 & & & & 0 & 2 & \\ \hline & & & & & & & 2 \end{array}\right]$$

compute

$$J-2I = \left[\begin{array}{ccc|cc|cc|c} 0 & 1 & 0 & & & & & \\ 0 & 0 & 1 & & & \multicolumn{3}{c}{0} \\ 0 & 0 & 0 & & & & & \\ \hline & & & 0 & 1 & & & \\ & & & 0 & 0 & & & \\ \hline & & & & & 0 & 1 & \\ & 0 & & & & 0 & 0 & \\ \hline & & & & & & & 0 \end{array}\right],$$

$$(J-2I)^2 = \left[\begin{array}{ccc|cc|cc|c} 0 & 0 & 1 & & & & & \\ 0 & 0 & 0 & & & \multicolumn{3}{c}{0} \\ 0 & 0 & 0 & & & & & \\ \hline & & & 0 & 0 & & & \\ & & & 0 & 0 & & & \\ \hline & & & & & 0 & 0 & \\ & 0 & & & & 0 & 0 & \\ \hline & & & & & & & 0 \end{array}\right]$$

and $(J-2I)^3=0$. Thus, we know that

$$\begin{aligned} (J-2I)^3 &= 0 \\ \operatorname{rank}(J-2I)^2 &= 1 \\ \operatorname{rank}(J-2I) &= 4 \\ (J-2I) &\text{ is 8-by-8} \end{aligned}$$

This list of numbers is sufficient to determine the block structure of J. The fact that $(J-2I)^3=0$ tells us that the largest block has order 3. The rank of $(J-2I)^2$ will be the number of blocks of order 3, so there is only one. The rank of $(J-2I)$ is twice the number of blocks of order 3 plus the number of blocks of order 2, so there are two of them. The number of blocks of order 1 is $8-(2\times 2)-3=1$. The same procedure can be applied to direct sums of Jordan blocks of any size so long as all the blocks correspond to the same eigenvalue. If J is such a direct sum corresponding to the eigenvalue λ, then the smallest integer k_1 such that $(J-\lambda I)^{k_1}=0$ is the size of the largest block. The rank of $(J-\lambda I)^{k_1-1}$ is the number of blocks of order k_1, the rank of $(J-\lambda I)^{k_1-2}$ is twice the number of blocks of order k_1 plus the number of blocks of size k_1-1, and so forth. The sequence of ranks of $(J-\lambda I)^{k_1-i}$, $i=0,1,2,\ldots,k_1-1$, recursively determines the orders of all the blocks in J.
6. The sizes of all the Jordan blocks in a general Jordan matrix (3.2.1.1) are determined by a knowledge of the ranks of certain powers. If λ_1 is an eigenvalue of a Jordan matrix $J\in M_n$, then only the Jordan blocks corresponding to λ_1 will be annihilated when one forms $(J-\lambda_1 I), (J-\lambda_1 I)^2, \ldots$ because the other blocks in $J-\lambda_1 I$ all have nonzero diagonal entries. Eventually, the rank of $(J-\lambda_1 I)^k$ will stop decreasing (one need not consider any $k>n$); the smallest value of k for which the rank of $(J-\lambda_1 I)^k$ attains its minimum value is the order of the largest block corresponding to λ_1. This minimum value is called the *index* of the eigenvalue λ_1. An analysis of the ranks of the sequence of powers of $J-\lambda_1 I$ is sufficient to determine the sizes and numbers of Jordan blocks corresponding to λ_1. One then proceeds to λ_2, λ_3, and so on and determines all the blocks in this same way.

Although the preceding list of observations was made about a Jordan matrix J, each observation also applies to any matrix that is similar to J. Thus, if one is given a matrix $A\in M_n$, the Jordan canonical form of A

(but not the similarity that transforms A to Jordan canonical form) can be determined by the following procedure:

1. Find all the distinct eigenvalues of A, perhaps by finding the roots of the characteristic polynomial.
2. For each distinct eigenvalue λ_i of A, form $(A-\lambda_i I)^k$ for $k=1,2,\ldots,n$ and analyze the sequence of ranks of these matrices to discover the orders and numbers of all the Jordan blocks of A corresponding to the eigenvalue λ_i.

This algorithm may be useful in analyzing small matrices of simple form by hand, but it is useless for automatic computation because determination of the rank of a matrix is an inherently unstable process. The example $A_\epsilon=\begin{bmatrix}1&0\\0&\epsilon\end{bmatrix}$ makes this clear; A_ϵ has rank 2 for all $\epsilon\neq 0$, but it has rank 1 for $\epsilon=0$.

As an example of a situation in which this algorithm is useful, consider the problem of determining the Jordan canonical form of the square of a Jordan block $J_k(0)\in M_k$:

$$A=\begin{bmatrix}0&1&&0\\&&\ddots&\\&&\ddots&1\\0&&&0\end{bmatrix}^2=\begin{bmatrix}0&0&1&\cdots&0\\&&\ddots&\ddots&\vdots\\&&&\ddots&1\\&&&\ddots&0\\0&&&&0\end{bmatrix}$$

The eigenvalues of A are all zero, $A^m=0$ for $m=[(k+1)/2]=$ the greatest integer in $(k+1)/2$, and $A^p\neq 0$ if $p=1,2,\ldots,m-1$. The rank of each power A^{p+1} is 2 less than the rank of its predecessor A^p for $p=1,2,\ldots,m-2$; A^{m-1} has rank 2 if k is even and has rank 1 if k is odd. Thus, the Jordan canonical form of $A=J_k^2(0)$ is

$$\begin{bmatrix}J_m(0)&0\\0&J_m(0)\end{bmatrix}\quad\text{if}\quad k=2m\text{ is even}$$

and

$$\begin{bmatrix}J_m(0)&0\\0&J_{m-1}(0)\end{bmatrix}\quad\text{if}\quad k=2m-1\text{ is odd}$$

This observation is useful in determining whether a given matrix has a square root. For example, it shows that there is no matrix $A\in M_2$ such that $A^2=\begin{bmatrix}0&1\\0&0\end{bmatrix}$.

3.2.2 Linear systems of ordinary differential equations. One application of the Jordan canonical form that is of considerable theoretical

importance is to the analysis of solutions of a system of ordinary differential equations with constant coefficients. Let $A\in M_n$ be given, and consider the first-order initial value problem

$$\begin{aligned}x'(t)&=Ax(t)\\x(0)&=x_0\text{ given}\end{aligned}\qquad(3.2.2.1)$$

where $x(t)=[x_1(t),x_2(t),\ldots,x_n(t)]^T$, and the prime (′) denotes differentiation with respect to t. If A is not a diagonal matrix, this system of equations is *coupled*; that is, $x_i'(t)$ is related not only to $x_i(t)$ but to the other components of the vector $x(t)$ as well. This coupling makes the problem hard to solve, but if A can be transformed to diagonal (or almost diagonal) form, the amount of coupling can be reduced or even eliminated and the problem may be easier to solve. If $A=SJS^{-1}$, where J is the Jordan canonical form of A, then (3.2.2.1) becomes

$$\begin{aligned}y'(t)&=Jy(t)\\y(0)&=y_0\text{ given}\end{aligned}\qquad(3.2.2.2)$$

where $x(t)=Sy(t)$ and $y_0=S^{-1}x_0$. If the problem (3.2.2.2) can be solved, then each component of the solution $x(t)$ to (3.2.2.1) is just a linear combination of the components of the solution to (3.2.2.2), and the linear combinations are given by S.

If A is diagonalizable, then J is a diagonal matrix, and (3.2.2.2) is just an uncoupled set of equations of the form $y_k'(t)=\lambda_k y_k(t)$, which have the solution $y_k(t)=y_k(0)e^{\lambda_k t}$. If the eigenvalue λ_k is real, this is a simple exponential, and if $\lambda_k=a_k+ib_k$ is complex, it is an oscillatory term $y_k(t)=y_k(0)e^{a_k t}[\cos(b_k t)+i\sin(b_k t)]$.

If J is not diagonal, the solution is more complicated. The components of $y(t)$ that correspond to different Jordan blocks in J are not coupled, so it suffices to consider the case in which J is a single Jordan block

$$J=\begin{bmatrix}\lambda&1&&0\\&&\ddots&\\&&\ddots&1\\0&&&\lambda\end{bmatrix}\in M_m$$

The system (3.2.2.2) is

$$\begin{aligned}y_1'(t)&=\lambda y_1(t)+y_2(t)\\&\ \ \vdots\\y_{m-1}'(t)&=\lambda y_{m-1}(t)+y_m(t)\\y_m'(t)&=\lambda y_m(t)\end{aligned}$$

which can be solved in a straightforward way from the bottom up. Starting with the last equation, we obtain

$$y_m(t) = y_m(0)e^{\lambda t}$$

so that

$$y'_{m-1}(t) = \lambda y_{m-1}(t) + y_m(0)e^{\lambda t}$$

This has the solution

$$y_{m-1}(t) = y_m(0)te^{\lambda t} + y_{m-1}(0)e^{\lambda t}$$

which can now be used in the next equation, which becomes

$$y'_{m-2}(t) = \lambda y_{m-2}(t) + y_m(0)te^{\lambda t} + y_{m-1}(0)e^{\lambda t}$$

This has the solution

$$y_{m-2}(t) = y_m(0)\frac{t^2}{2}e^{\lambda t} + y_{m-1}(0)te^{\lambda t} + y_{m-2}(0)e^{\lambda t}$$

and so forth. It is clear that each component of the solution is of the form

$$y_k(t) = e^{\lambda t}q_k(t)$$

where $q_k(t)$ is a polynomial of degree at most $m-k$, $k = 1, \ldots, m$.

From this analysis, we conclude that for any given initial condition x_0, the components of the solution $x(t)$ of the problem (3.2.2.1) have the form

$$x_j(t) = p_1(t)e^{\lambda_1 t} + p_2(t)e^{\lambda_2 t} + \cdots + p_k(t)e^{\lambda_k t}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k$ are the distinct eigenvalues of A and each $p_i(t)$ is a polynomial whose degree is strictly less than the order of the largest Jordan block corresponding to the eigenvalue λ_i. Real eigenvalues produce pure exponential terms, while complex eigenvalues may produce mixed exponential and oscillatory terms.

3.2.3 **Similarity of a matrix and its transpose.** Every Jordan block is permutation-similar to its transpose, as can be seen from the similarity

$$\begin{bmatrix} 0 & & 1 \\ & ⋰ & \\ 1 & & 0 \end{bmatrix} \begin{bmatrix} \lambda & 1 & & 0 \\ & \ddots & \ddots & \\ & & & 1 \\ 0 & & & \lambda \end{bmatrix} \begin{bmatrix} 0 & & 1 \\ & ⋰ & \\ 1 & & 0 \end{bmatrix} = \begin{bmatrix} \lambda & & & 0 \\ 1 & \ddots & & \\ & \ddots & & \\ 0 & & 1 & \lambda \end{bmatrix}$$

Therefore, if $A \in M_n$ is given and $A = SJS^{-1}$ is its Jordan canonical form, then A is similar to J, which is similar to J^T, which is similar to $A^T = (S^T)^{-1}J^T(S^T)$. The conclusion is that every complex matrix is similar to its transpose. From this it follows that the row rank (the maximum number of linearly independent rows) of a complex matrix is the same as its column rank (the maximum number of linearly independent columns) because rank is a similarity invariant. It also follows from this that A and A^T have the same eigenvalues, but all of these consequences are more easily established directly.

It is also the case that every matrix in $M_n(\mathbf{F})$ is similar, via a matrix in $M_n(\mathbf{F})$, to its transpose for any field $\mathbf{F}$; it is not necessary to assume that $\mathbf{F} = \mathbf{C}$. In fact, the similarity matrix may be taken to be symmetric.

3.2.4 **Commuting matrices and nonderogatory matrices.** If $p(t)$ is a polynomial, and if $A \in M_n$ is a given matrix, it is a useful, if obvious, fact that $p(A)$ commutes with A. What about the converse? If $A, B \in M_n$ are given and if A commutes with B, is B necessarily a polynomial in A? Evidently not, for if we take $A = I$, then A commutes with every matrix and $p(I) = p(1)I$ cannot produce any nonscalar matrix. The problem is that the form of A permits it to commute with many matrices, but permits it to generate only a few matrices of the form $p(A)$. To get any result, some compromise between these two forces must be found.

3.2.4.1 **Definition.** A matrix $A \in M_n$ is said to be *nonderogatory* if every eigenvalue of A has geometric multiplicity 1.

Since the geometric multiplicity of an eigenvalue of a Jordan matrix is equal to the number of Jordan blocks corresponding to that eigenvalue, a matrix is nonderogatory if and only if corresponding to each distinct eigenvalue is exactly one Jordan block. A matrix $A \in M_n$ is nonderogatory, for example, if it has n distinct eigenvalues or if it has only one eigenvalue and that eigenvalue has geometric multiplicity 1. A scalar matrix is the antithesis of a nonderogatory matrix.

3.2.4.2 **Theorem.** Let $A \in M_n$ be a given nonderogatory matrix. A matrix $B \in M_n$ commutes with A if and only if there is a polynomial $p(\cdot)$ of degree at most $n-1$ such that $B = p(A)$.

Proof: If $B = p(A)$, then certainly A commutes with B. For the converse, let $A = SJS^{-1}$ be the Jordan canonical form of A. If $BA = AB$, then $BSJS^{-1} = SJS^{-1}B$ and $(S^{-1}BS)J = J(S^{-1}BS)$. If we can show that $S^{-1}BS = p(J)$, then $B = Sp(J)S^{-1} = p(SJS^{-1}) = p(A)$. Thus, it suffices

to assume that A is itself a Jordan matrix. Since A is nonderogatory, we may assume that

$$A = \begin{bmatrix} J_{n_1}(\lambda_1) & & & 0 \\ & J_{n_2}(\lambda_2) & & \\ & & \ddots & \\ 0 & & & J_{n_k}(\lambda_k) \end{bmatrix}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_k$ are the k distinct eigenvalues of A. If we write B in partitioned form $B=(B_{ij})$, which conforms with this decomposition of A, then the corresponding off-diagonal blocks of $AB-BA$ are of the form

$$J_{n_i}(\lambda_i)B_{ij} - B_{ij}J_{n_j}(\lambda_j) = 0, \qquad i \neq j$$

Since the eigenvalues λ_i and λ_j are different, we can conclude [see Problem 9 in Section (2.4)] that $B_{ij}=0$ is the unique solution of these equations. The matrix B must therefore be a block diagonal matrix

$$B = \begin{bmatrix} B_1 & & & 0 \\ & B_2 & & \\ & & \ddots & \\ 0 & & & B_k \end{bmatrix}$$

with each $B_i \in M_{n_i}$. The commutativity assumption says that $B_iJ_{n_i}(\lambda_i) = J_{n_i}(\lambda_i)B_i$ for all $i=1,2,\ldots,k$. If we write $J_{n_i}(\lambda_i) = \lambda_i I + N_i$ with

$$N_i = \begin{bmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{bmatrix} \in M_{n_i}$$

these identities become $B_iN_i = N_iB_i$ for $i=1,2,\ldots,k$. An explicit calculation shows that, because of the special form of N_i, each B_i must be an upper triangular matrix of Toeplitz type (0.9.7), that is,

$$B_i = \begin{bmatrix} b_1^{(i)} & b_2^{(i)} & \cdots & b_{n_i}^{(i)} \\ & b_1^{(i)} & \ddots & \vdots \\ & & \ddots & b_2^{(i)} \\ 0 & & & b_1^{(i)} \end{bmatrix} \qquad (3.2.4.3)$$

where the entries are constant down the diagonals. If we can construct polynomials $p_i(t)$ of degree at most $n-1$ with the property that

$p_i(J_{n_j}(\lambda_j))=0$ for all $i \neq j$, and $p_i(J_{n_i}(\lambda_i)) = B_i$, then

$$p(t) = p_1(t) + \cdots + p_k(t)$$

will fulfill the assertions of the theorem. Define

$$q_i(t) = \prod_{\substack{j=1 \\ j \neq i}}^{k} (t-\lambda_j)^{n_j}, \qquad \text{degree } q_i(t) = n - n_i$$

and observe that $q_i(J_{n_j}(\lambda_j))=0$ for all $i \neq j$ because $(J_{n_j}(\lambda_j)-\lambda_j I)^{n_j}=0$. Although $q_i(J_{n_i}(\lambda_i))$ is not necessarily equal to B_i, it is nonsingular (because the λ_i are distinct) and, like any polynomial in $J_{n_i}(\lambda_i)$, is of the form (3.2.4.3).

Since the inverse of any nonsingular matrix of the form (3.2.4.3) is of the same form, and since the product of any two matrices of this type is of the same type, the matrix

$$[q_i(J_{n_i}(\lambda_i))]^{-1}B_i$$

is an upper triangular matrix of Toeplitz type (3.2.4.3). Every such matrix can be written as a polynomial in $J_{n_i}(\lambda_i)$, for example,

$$B_i = b_1^{(i)}(J_{n_i}(\lambda_i)-\lambda_i I)^0 + b_2^{(i)}(J_{n_i}(\lambda_i)-\lambda_i I)^1 + \cdots + b_{n_i}^{(i)}(J_{n_i}(\lambda_i)-\lambda_i I)^{n_i-1}$$

Thus, there is a polynomial $r_i(t)$ of degree at most n_i-1 such that

$$[q_i(J_{n_i}(\lambda_i))]^{-1}B_i = r_i(J_{n_i}(\lambda_i))$$

If we now set $p_i(t) = q_i(t)r_i(t)$, the degree of $p_i(t)$ is at most $n-1$,

$$p_i(J_{n_j}(\lambda_j)) = q_i(J_{n_j}(\lambda_j))r_i(J_{n_j}(\lambda_j)) = 0r_i(J_{n_j}(\lambda_j)) = 0$$

if $i \neq j$, and $p_i(J_{n_i}(\lambda_i)) = q_i(J_{n_i}(\lambda_i))r_i(J_{n_i}(\lambda_i)) = B_i$. □

The converse of the theorem is also true, and leads to a characterization of nonderogatory matrices: A matrix $A \in M_n$ is nonderogatory if and only if every matrix that commutes with A is a polynomial in A.

3.2.5 **Convergent matrices.** A matrix $A \in M_n$ with the property that all elements of A^m tend to zero as $m \to \infty$ is called a *convergent* matrix. Such matrices play an important role in the analysis of algorithms in numerical linear algebra. If A is a diagonal matrix, then it is apparent that A is convergent if and only if all the eigenvalues of A are of modulus strictly less than 1, and the same reasoning extends to diagonalizable matrices.

Because of the limiting operation involved, it is not clear how to use a perturbation argument to extend this result to the general case of not necessarily diagonalizable matrices. We may use the Jordan canonical

form, however. If $A = SJS^{-1}$ is the Jordan canonical form of A, then $A^m = SJ^mS^{-1}$, so $A^m \to 0$ as $m \to \infty$ if and only if $J^m \to 0$ as $m \to \infty$. Since J is a direct sum of Jordan blocks, it suffices to consider the behavior of powers of a Jordan block

$$J_k(\lambda) = \begin{bmatrix} \lambda & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{bmatrix} = \begin{bmatrix} \lambda & & 0 \\ & \ddots & \\ 0 & & \lambda \end{bmatrix} + \begin{bmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{bmatrix}$$

$$= \lambda I + N_k \in M_k, \quad \text{where } N_k \equiv J_k(0)$$

Since $N_k^m = 0$ for all $m \ge k$, we have

$$[J_k(\lambda)]^m = (\lambda I + N_k)^m = \sum_{i=0}^{m} \binom{m}{i} \lambda^i N_k^{m-i} = \sum_{i=m-k+1}^{m} \binom{m}{i} \lambda^i N_k^{m-i}$$

for all $m \ge k$. Since the diagonal elements are all λ^m, if $J^m \to 0$ it is necessary that $\lambda^m \to 0$, which means that $|\lambda| < 1$. Conversely, if $|\lambda| < 1$, we would like to prove that

$$\binom{m}{m-j} \lambda^{m-j} \to 0 \quad \text{as} \quad m \to \infty \quad \text{for each} \quad j = 0, 1, 2, \ldots, k-1$$

But

$$\left| \binom{m}{m-j} \lambda^{m-j} \right| = \left| \frac{m(m-1)(m-2)\cdots(m-j+1)\lambda^m}{j!\lambda^j} \right| \le \left| \frac{m^j \lambda^m}{j!\lambda^j} \right|$$

so it will suffice to show that $m^j |\lambda|^m \to 0$ as $m \to \infty$. An easy way to see this is to take logarithms and observe that

$$j \log m + m \log|\lambda| \to -\infty$$

as $m \to \infty$ because $\log|\lambda| < 0$ and $(\log x)/x \to 0$ as $x \to \infty$ by l'Hopital's rule.

This argument has made essential use of the Jordan canonical form of A to show that $A^m \to 0$ as $m \to \infty$ if and only if all the eigenvalues of A have modulus less than 1. Another proof, which is completely independent of the Jordan canonical form, is given in (5.6.12).

3.2.6 **The geometric multiplicity–algebraic multiplicity inequality.** The geometric multiplicity of an eigenvalue of a given matrix $A \in M_n$ is the number of Jordan blocks of A corresponding to the eigenvalue. This number is less than or equal to the sum of the orders of all the Jordan blocks corresponding to the eigenvalue. This sum is the algebraic multiplicity. Thus, the geometric multiplicity of an eigenvalue is not greater than its algebraic multiplicity; compare with (1.4.9).

3.2.7 **Diagonalizable and nilpotent matrices.** A matrix $A \in M_n$ is *nilpotent* if $A^k = 0$ for some positive integer k. Any Jordan block $J_k(\lambda)$ can be written as $J_k(\lambda) = \lambda I + N_k$, where $(N_k)^k = 0$. Thus, any Jordan block is the sum of a diagonal matrix and a nilpotent matrix.

More generally, a Jordan matrix (3.2.1.1) can be written as $J = D + N$, where D is a diagonal matrix whose main diagonal is the same as that of J, and $N = J - D$. The matrix N is nilpotent, and $N^k = 0$ if k is the order of the largest Jordan block in J.

Finally, if $A \in M_n$ is any given matrix and $A = SJS^{-1}$ is in Jordan canonical form, then $A = SDS^{-1} + SNS^{-1} \equiv A_D + A_N$, where A_D is diagonalizable, A_N is nilpotent, and $A_D A_N = A_N A_D$ because both D and N are block diagonal matrices with respective blocks of the same size, and the blocks in D are scalar matrices.

We conclude that any $A \in M_n$ can be written as the sum of a diagonalizable matrix and a nilpotent matrix in such a way that these summands commute.

Problems

1. Let $\mathfrak{F} = \{A_\alpha : \alpha \in \mathfrak{I}\} \subset M_n$ be a given family of matrices, indexed by the index set $\mathfrak{I}$, and suppose there is a nonderogatory matrix $A_0 \in \mathfrak{F}$ such that $A_\alpha A_0 = A_0 A_\alpha$ for all $\alpha \in \mathfrak{I}$. Show that for every $\alpha \in \mathfrak{I}$ there is a polynomial $p_\alpha(t)$ of degree at most $n-1$ such that $A_\alpha = p_\alpha(A_0)$, and hence $\mathfrak{F}$ is a commuting family.

2. Let $A \in M_n$ be given, and let λ_i be an eigenvalue of A. Show that the order of the largest Jordan block of A corresponding to the eigenvalue λ_i (the *index* of λ_i) is the smallest value of $k = 1, 2, \ldots, n-1$ for which $\operatorname{rank}(A - \lambda_i I)^k = \operatorname{rank}(A - \lambda_i I)^{k+1}$.

3. If $A \in M_n$ has $A^k = 0$ for some $k > n$, show that $A^r = 0$ for some $r \le n$. Thus, every nilpotent matrix has a vanishing power that is not greater than the order of the matrix. *Hint:* Show that 0 is the only eigenvalue of A. What does the Jordan canonical form of A look like? Now take powers.

4. Let $J_k(0)$ be a given Jordan block. Use the argument at the end of (3.2.1) to determine the three possible Jordan canonical forms of the matrix $J_k^3(0)$.

5. Let $A \in M_n$ be nilpotent, so $A^k = 0$ for some k. Show that the characteristic polynomial of A is $p_A(t) = t^n$.

6. The linear transformation $d/dt: p(t) \to p'(t)$ acting on the vector space of all polynomials with degree at most 3 has the basis representation

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

in the basis $B = \{1, t, t^2, t^3\}$. What is the Jordan canonical form of this matrix?

7. What are the possible Jordan forms of a matrix $A \in M_n$ such that $A^3 = I$?

8. What are the possible Jordan canonical forms for a matrix $A \in M_6$ with characteristic polynomial $p_A(t) = (t+3)^4(t-4)^2$?

9. Use the method described in (3.2.1) to determine the Jordan canonical form of

$$A = \begin{bmatrix} i & 1 \\ 1 & -i \end{bmatrix}$$

10. Verify the assertion in the proof of Theorem (3.2.4.2) that products of matrices of the form (3.2.4.3) have the same form. Deduce that the inverse of a nonsingular matrix of the form (3.2.4.3) has the same form. *Hint:* The inverse of A is a polynomial in A.

11. Let $A, B \in M_n$. Use the identities in the proof of Theorem (1.3.20) to show that the nonsingular blocks in the Jordan forms of AB and BA are identical. Does this mean that AB and BA are similar? If AB and BA are not similar, discuss how far from similar they can be.

12. Suppose that $A_1, \ldots, A_k$ are given matrices with $A_i \in M_{n_i}$ for $i = 1, 2, \ldots, k$, and suppose that $J_1, \ldots, J_k$ are their respective Jordan canonical forms. Show that the Jordan canonical form of the direct sum

$$\begin{bmatrix} A_1 & & & 0 \\ & A_2 & & \\ & & \ddots & \\ 0 & & & A_k \end{bmatrix} \in M_n, \qquad n_1 + n_2 + \cdots + n_k = n$$

has (up to permutation of diagonal subblocks) the Jordan canonical form

$$\begin{bmatrix} J_1 & & & 0 \\ & J_2 & & \\ & & \ddots & \\ 0 & & & J_k \end{bmatrix}$$

13. Let $A \in M_n$ and $B, C \in M_m$ be given. Show that the direct sum $\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix} \in M_{n+m}$ is similar to the direct sum $\begin{bmatrix} A & 0 \\ 0 & C \end{bmatrix}$ if and only if B is similar to C.

14. Let $B, C \in M_m$ be given. Show that the two k-fold direct sums

$$\begin{bmatrix} B & & & 0 \\ & B & & \\ & & \ddots & \\ 0 & & & B \end{bmatrix} \in M_{km} \quad \text{and} \quad \begin{bmatrix} C & & & 0 \\ & C & & \\ & & \ddots & \\ 0 & & & C \end{bmatrix} \in M_{km}, \qquad k \geq 1$$

are similar if and only if B and C are similar.

15. Let $A \in M_n$ and $B, C \in M_m$ be given. Show that the direct sums

$$\begin{bmatrix} A & & & 0 \\ & B & & \\ & & \ddots & \\ 0 & & & B \end{bmatrix} \in M_{n+km} \quad \text{and} \quad \begin{bmatrix} A & & & 0 \\ & C & & \\ & & \ddots & \\ 0 & & & C \end{bmatrix} \in M_{n+km},$$

$$k \geq 1$$

are similar if and only if B and C are similar.

16. Let $A \in M_n$ have Jordan canonical form $J_{n_1}(\lambda_1) \oplus \cdots \oplus J_{n_k}(\lambda_k)$. If A is nonsingular, show that the Jordan canonical form of A^2 is $J_{n_1}(\lambda_1^2) \oplus \cdots \oplus J_{n_k}(\lambda_k^2)$; that is, the Jordan canonical form of A^2 is composed of precisely the same collection of Jordan blocks as A, but the respective eigenvalues are squared. Is a statement like this true for all powers A^m, $m \geq 2$? Give a 2-by-2 example to show this is false if A is singular. *Hint:* If $\lambda \neq 0$, show that the Jordan canonical form of $J_k^2(\lambda)$ (a simple Jordan block) is $J_k(\lambda^2)$. Show that

$$\operatorname{rank}[J_k(\lambda) - \lambda I]^m = \operatorname{rank}[J_k^2(\lambda) - \lambda^2 I]^m, \qquad m = 1, 2, \ldots, k, \quad \text{if} \quad \lambda \neq 0$$

17. If $A \in M_n$, show that $\operatorname{rank} A = \operatorname{rank} A^2$ if and only if the geometric and algebraic multiplicities of the eigenvalue $\lambda = 0$ are equal; that is, all

the Jordan blocks corresponding to $\lambda=0$ (if any) in the Jordan canonical form of A are 1-by-1.

Further Readings. For a proof of the fact that the similarity matrix between a given matrix and its transpose may always be taken to be symmetric, see 0. Taussky and H. Zassenhaus, "On the Similarity Transformation between a Matrix and Its Transpose," *Pacific J. Math.* 9 (1959), 893–896. See [HJ] for a proof of the converse to Theorem (3.2.4.2) mentioned at the end of Section (3.2.4).

3.3 Polynomials and matrices: the minimal polynomial

If $p(t)=t^k+a_{k-1}t^{k-1}+a_{k-2}t^{k-2}+\cdots+a_1t+a_0$ is a given polynomial, then one can always define

$$p(A)\equiv A^k+a_{k-1}A^{k-1}+a_{k-2}A^{k-2}+\cdots+a_1A+a_0I$$

for any $A\in M_n$. There is an important interplay between polynomials and matrices. The vital role of the characteristic polynomial has already been observed, but there are other polynomials associated with a square matrix. One of these is the minimal polynomial.

The Cayley–Hamilton theorem (2.4.2) guarantees that for each $A\in M_n$ there is a polynomial (the characteristic polynomial) $p_A(t)$ of degree n such that $p_A(A)=0$. A polynomial whose value is the 0 matrix at A is said to *annihilate* A. There may also be a polynomial of degree $n-1$ which annihilates A, or one of degree $n-2$, but it is clear that, since there are only finitely many possibilities, for each $A\in M_n$ there is a polynomial of minimum degree that annihilates A, and this minimum degree is at most n. If $p(A)=0$, then $cp(A)=0$ for any $c\in \mathbf{C}$, so it is clear that we may always normalize a nontrivial annihilating polynomial so that the coefficient of the highest-order term is $+1$. A polynomial whose highest-order term has coefficient $+1$ is said to be *monic*. Notice that a monic polynomial cannot be identically zero.

3.3.1 Theorem. Let $A\in M_n$ be given. There exists a unique monic polynomial $q_A(t)$ of minimum degree that annihilates A. The degree of this polynomial is at most n. If $p(t)$ is any polynomial such that $p(A)=0$, then $q_A(t)$ divides $p(t)$.

Proof: The characteristic polynomial is an example of a polynomial of degree n that annihilates A, so there is a minimum positive integer $m\le n$ such that there exists a monic polynomial $q(t)$ of degree m with $q(A)=0$.

If $p(t)$ annihilates A, and if $q(t)$ is a monic polynomial of minimum degree that annihilates A, then the degree of $q(t)$ must be less than or equal to the degree of $p(t)$. By the Euclidean algorithm, therefore, there exists a polynomial $h(t)$ and a polynomial $r(t)$ of degree less than that of $q(t)$ such that $p(t)=q(t)h(t)+r(t)$. But $0=p(A)=q(A)h(A)+r(A)=0h(A)+r(A)$, so $r(A)=0$. If $r(t)\not\equiv 0$, we could normalize it and obtain a monic polynomial of degree less than that of $q(t)$ that annihilates A. Since this would contradict the minimal property of $q(t)$, we conclude that $r(t)\equiv 0$, and hence that $q(t)$ divides $p(t)$ with quotient $h(t)$. If there are two monic polynomials of minimum degree that annihilate A, this argument shows that each divides the other; since the degrees are the same, one must be a scalar multiple of the other. But since both are monic, the scalar factor must be $+1$ and they are identical. □

3.3.2 Definition. Let $A\in M_n$ be given. The unique monic polynomial $q_A(t)$ of minimum degree that annihilates A is called the *minimal polynomial* of A.

3.3.3 Corollary. Similar matrices have the same minimal polynomial.

Proof: If $A,B,S\in M_n$ and if $A=SBS^{-1}$, then $q_B(A)=q_B(SBS^{-1})=Sq_B(B)S^{-1}=0$, so the degree of $q_B(t)$ is not less than the degree of $q_A(t)$. But $B=S^{-1}AS$, so the same argument shows that the degree of $q_A(t)$ is not less than the degree of $q_B(t)$. Thus, these two monic polynomials have the same minimal degree and both annihilate A, so they must be identical by Theorem (3.3.1). □

3.3.4 Corollary. For every $A\in M_n$, the minimal polynomial $q_A(t)$ divides the characteristic polynomial $p_A(t)$. Moreover, $q_A(\lambda)=0$ if and only if λ is an eigenvalue of A, so every root of $p_A(t)=0$ is a root of $q_A(t)=0$.

Proof: Since $p_A(A)=0$, the fact that there is a polynomial $h(t)$ such that $p_A(t)=h(t)q_A(t)$ follows from the theorem. This factorization makes it clear that every root of $q_A(t)=0$ is a root of $p_A(t)=0$, and hence every root of $q_A(t)=0$ is an eigenvalue of A. If λ is an eigenvalue of A, and if $x\neq 0$ is an associated eigenvector, then $Ax=\lambda x$ and $0=q_A(A)x=q_A(\lambda)x$, so $q_A(\lambda)=0$. □

This last corollary shows that if the characteristic polynomial $p_A(t)$ has been completely factored as

$$p_A(t) = \prod_{i=1}^{m} (t-\lambda_i)^{s_i}, \qquad 1 \le s_i \le n, \qquad s_1+s_2+\cdots+s_m = n \tag{3.3.5a}$$

with $\lambda_1, \lambda_2, \ldots, \lambda_m$ distinct, then the minimal polynomial $q_A(t)$ must have the form

$$q_A(t) = \prod_{i=1}^{m} (t-\lambda_i)^{r_i}, \qquad 1 \le r_i \le s_i \tag{3.3.5b}$$

In principle, this gives an algorithm for finding the minimal polynomial of a given matrix A:

1. First compute the eigenvalues of A, together with their algebraic multiplicities, perhaps by finding the characteristic polynomial and factoring it completely. By some means, determine the factorization (3.3.5a).
2. There are finitely many polynomials of the form of the product in (3.3.5b). Starting with the product in which all $r_i = 1$, determine by explicit calculation the one of minimal degree that annihilates A. This will be the minimal polynomial.

Numerically, this is not a good algorithm if it involves factoring the characteristic polynomial of a large matrix, but it can be very effective for hand calculations involving small matrices of simple form. Another approach to computing the minimal polynomial that does not involve knowing either the characteristic polynomial or the eigenvalues is outlined in Problem 5 at the end of this section.

There is an intimate connection between the Jordan canonical form of A and the minimal polynomial of A. Suppose $A = SJS^{-1}$ is the Jordan canonical form of A, and suppose first that

$$J = \begin{bmatrix} \lambda & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{bmatrix} \in M_n$$

is only a single Jordan block. The characteristic polynomial of A is $(t-\lambda)^n$, and since $(J-\lambda I)^k \ne 0$ if $k < n$, the minimal polynomial is also $(t-\lambda)^n$. If

$$J = \begin{bmatrix} J_{n_1}(\lambda) & 0 \\ 0 & J_{n_2}(\lambda) \end{bmatrix} \in M_n$$

with $n_1 \ge n_2$, then the characteristic polynomial of J is still $(t-\lambda)^n$, but now $(J-\lambda I)^{n_1} = 0$ and no lower power vanishes. The minimal polynomial is therefore $(t-\lambda)^{n_1}$. If there are more blocks, the result is the

same: The minimal polynomial of J is $(t-\lambda)^r$, where r is the order of the largest Jordan block corresponding to λ. If J is a general Jordan matrix, the minimal polynomial must contain a factor $(t-\lambda_i)^{r_i}$ for each distinct eigenvalue λ_i, and r_i must be the order of the largest Jordan block corresponding to λ_i; no smaller power will annihilate all the Jordan blocks corresponding to λ_i, and no greater power is needed. Since similar matrices have the same minimal polynomial, we have proved the following theorem.

3.3.6 **Theorem.** Let $A \in M_n$ be a given matrix whose distinct eigenvalues are $\lambda_1, \lambda_2, \ldots, \lambda_m$. The minimal polynomial of A is

$$q_A(t) = \prod_{i=1}^{m} (t-\lambda_i)^{r_i} \tag{3.3.7}$$

where r_i is the order of the largest Jordan block of A corresponding to the eigenvalue λ_i.

In practice, this result is not very helpful in computing the minimal polynomial since it is usually harder to determine the Jordan canonical form of a matrix than it is to determine its minimal polynomial. After all, if only the eigenvalues of a matrix are known, its minimal polynomial can be determined by simple trial and error. There are important theoretical consequences, however. Since a matrix is diagonalizable if and only if all its Jordan blocks have order 1, a necessary and sufficient condition for diagonalizability is that all $r_i = 1$ in (3.3.7).

3.3.8 **Corollary.** Let $A \in M_n$ be a given matrix whose distinct eigenvalues are $\lambda_1, \lambda_2, \ldots, \lambda_m$. Then A is diagonalizable if and only if $q(A) = 0$, where

$$q(t) = (t-\lambda_1)(t-\lambda_2)\cdots(t-\lambda_m) \tag{3.3.9}$$

This criterion is actually useful for determining if a given matrix is diagonalizable, for if one knows the eigenvalues of a given matrix, it is easy to form the polynomial (3.3.9) and see if it annihilates A. If it does, it must be the minimal polynomial of A, since no lower-order polynomial could have as roots all the m distinct eigenvalues of A. It is sometimes useful to have this result formulated in several equivalent ways:

3.3.10 **Corollary.** Let $A \in M_n$ be given. Each of the following is a necessary and sufficient condition for A to be diagonalizable:

(a) The minimal polynomial $q_A(t)$ has distinct linear factors.
(b) Every root of $q_A(t) = 0$ has multiplicity 1.
(c) For all t such that $q_A(t) = 0$, the derivative $q'_A(t) \ne 0$.

We have been considering the problem of finding, for a given matrix $A \in M_n$, a monic polynomial of minimum degree that annihilates A. But what about the converse? Given a monic polynomial

$$p(t) = t^n + a_{n-1}t^{n-1} + a_{n-2}t^{n-2} + \cdots + a_1 t + a_0 \tag{3.3.11}$$

is there a matrix A for which $p(t)$ is the minimal polynomial? If so, the size of A must be at least n-by-n; it is not hard to find such a matrix $A \in M_n$. Consider the matrix

$$A = \begin{bmatrix} 0 & & & & -a_0 \\ 1 & 0 & & 0 & \vdots \\ & 1 & \ddots & & \vdots \\ & & \ddots & 0 & -a_{n-2} \\ & 0 & & 1 & -a_{n-1} \end{bmatrix} \in M_n \tag{3.3.12}$$

and observe that

$$\begin{aligned} Ie_1 &= e_1 = A^0 e_1 \\ Ae_1 &= e_2 = Ae_1 \\ Ae_2 &= e_3 = A^2 e_1 \\ Ae_3 &= e_4 = A^3 e_1 \\ &\vdots \\ Ae_{n-1} &= e_n = A^{n-1} e_1 \\ Ae_n &= -a_{n-1}e_n - a_{n-2}e_{n-1} - \cdots - a_1 e_2 - a_0 e_1 \\ &= -a_{n-1}A^{n-1}e_1 - a_{n-2}A^{n-2}e_1 - \cdots - a_1 A e_1 - a_0 e_1 = A^n e_1 \\ &= [A^n - p(A)]e_1 \end{aligned}$$

Thus,

$$\begin{aligned} p(A)e_1 &= (a_0 e_1 + a_1 A e_1 + a_2 A^2 e_1 + \cdots + a_{n-1}A^{n-1}e_1) + A^n e_1 \\ &= [p(A) - A^n]e_1 + [A^n - p(A)]e_1 = 0 \end{aligned}$$

Furthermore, $p(A)e_k = p(A)A^{k-1}e_1 = A^{k-1}p(A)e_1 = A^{k-1}0 = 0$ for each $k = 1, 2, \ldots, n$. Since $p(A)e_k = 0$ for every basis vector e_k, we conclude that $p(A) = 0$. Thus $p(t)$ is a monic polynomial of degree n that annihilates A. If there were a polynomial $q(t) = t^m + b_{m-1}t^{m-1} + \cdots + b_1 t + b_0$ of lower degree $m < n$ that annihilates A, then

$$\begin{aligned} 0 = q(A)e_1 &= A^m e_1 + b_{m-1}A^{m-1}e_1 + \cdots + b_1 A e_1 + b_0 e_1 \\ &= e_{m+1} + b_{m-1}e_m + \cdots + b_1 e_2 + b_0 e_1 = 0 \end{aligned}$$

which would imply that the basis vector e_{m+1} is linearly dependent on the basis vectors $e_1, e_2, \ldots, e_m$. Since this is impossible, we conclude that $p(t)$ is the unique monic polynomial of minimum order that annihilates A. Moreover, since $p(t)$ has degree n, $A \in M_n$, and the characteristic polynomial $p_A(t)$ is a monic polynomial of degree n that also annihilates A, (3.3.11) must be the characteristic polynomial of (3.3.12).

3.3.13 **Definition.** The matrix (3.3.12) is known as the *companion matrix* of the polynomial (3.3.11).

We have proved the following:

3.3.14 **Theorem.** Every monic polynomial is both the minimal polynomial and the characteristic polynomial of its companion matrix.

Later, we shall develop methods to determine regions that contain the eigenvalues of a matrix. Since the zeroes of a polynomial are the eigenvalues of its companion matrix, these methods can be used to locate the zeroes of a polynomial. See Section (5.6).

If $A \in M_n$ is a given matrix, one can compute the characteristic polynomial $p_A(t)$ and the companion matrix (3.3.12) of the polynomial $p_A(t)$. If A is similar to this companion matrix, then (since similar matrices have the same minimal polynomial) it follows from (3.3.14) that the minimal polynomial $q_A(t)$ of A must be identical to the characteristic polynomial $p_A(t)$. This will not be the case in general, but if $A \in M_n$ is a matrix whose minimal polynomial $q_A(t)$ and characteristic polynomial $p_A(t)$ are identical, then the Jordan canonical form (3.1.12) of A must contain exactly one Jordan block for each distinct eigenvalue. The size of each Jordan block is equal to the multiplicity of the corresponding eigenvalue as a zero of the characteristic (minimal) polynomial of A. But the Jordan canonical form of the companion matrix of the polynomial $p_A(t)$ has exactly the same Jordan block structure, and hence it must be similar to A. This argument is a proof of the following:

3.3.15 **Theorem.** A matrix $A \in M_n$ is similar to the companion matrix of its characteristic polynomial if and only if the minimal and characteristic polynomials of A are identical.

Exercise. Show that $A \in M_n$ is similar to the companion matrix of its characteristic polynomial if and only if A is nonderogatory.

Problems

1. Let $A, B \in M_3$ be nilpotent. Show that A and B are similar if and only if A and B have the same minimal polynomial. Is this true in M_4?

2. Suppose $A \in M_n$ is given and suppose the distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_m$ of A are known. Use (3.3.6) to show that the minimal polynomial (3.3.7) is determined by the following algorithm: For each $i = 1, 2, \ldots, m$ compute $(A-\lambda_i I)^k$ for $k = 1, 2, \ldots, n$. Let r_i be the smallest value of k for which $\operatorname{rank}(A-\lambda_i I)^k = \operatorname{rank}(A-\lambda_i I)^{k+1}$. This number r_i is known as the *index* of the eigenvalue λ_i.

3. A matrix $A \in M_n$ is *idempotent* if $A^2 = A$. Use (3.3.10) to show that every idempotent matrix is diagonalizable. *Hint:* Show that $t^2 - t = t(t-1)$ annihilates A. What is the minimal polynomial of A? What can you say if A is *tripotent* ($A^3 = A$)? What if $A^k = A$?

4. If $A \in M_n$ has $A^k = 0$ for some $k > n$, show that $A^r = 0$ for some $r \leq n$. Thus, every nilpotent matrix has a vanishing power that is not greater than the order of the matrix. *Hint:* If $p(t) = t^k$ annihilates A, what does (3.3.1) say about the minimal polynomial?

5. Show that the following application of the Gram–Schmidt process permits one to compute directly the minimal polynomial of a given matrix $A \in M_n$ without knowing either the characteristic polynomial of A or any of its eigenvalues.

(a) Let the mapping $T: M_n \to \mathbf{C}^{n^2}$ be defined as follows: For any $A \in M_n$ partitioned according to columns as $A = [a_1\ a_2 \ldots a_n]$, let $T(A)$ denote the unique vector in $\mathbf{C}^{n^2}$ whose first n entries are the entries of the first column a_1, whose entries from $n+1$ to $2n$ are the entries of the second column a_2, and so forth. Show that this mapping T is an isomorphism (linear, one-to-one, and onto) of the vector spaces M_n and $\mathbf{C}^{n^2}$.

(b) Consider the vectors

$$v_0 = T(I),\ v_1 = T(A),\ v_2 = T(A^2), \ldots, v_k = T(A^k), \ldots$$

in $\mathbf{C}^{n^2}$ for $k = 0, 1, 2, \ldots, n$. Use the Cayley–Hamilton theorem to show that $\{v_0, v_1, \ldots, v_n\}$ is a dependent set.

(c) Apply the Gram–Schmidt process to the set $\{v_0, v_1, \ldots, v_n\}$ in the given order until it stops by producing a first zero vector. Why must a zero vector be produced?

(d) If the Gram–Schmidt process produces a first zero vector at the kth step, argue that $k-1$ is the degree of the minimal polynomial of A.

(e) If the kth step of the Gram–Schmidt process produces the vector $\alpha_0 v_0 + \alpha_1 v_1 + \cdots + \alpha_{k-1} v_{k-1} = 0$, show that

$$T^{-1}(\alpha_0 v_0 + \alpha_1 v_1 + \cdots + \alpha_{k-1} v_{k-1}) = \alpha_0 I + \alpha_1 A + \alpha_2 A^2 + \cdots + \alpha_{k-1} A^{k-1} = 0$$

and conclude that $q_A(t) = (\alpha_{k-1} t^{k-1} + \cdots + \alpha_2 t^2 + \alpha_1 t + \alpha_0)/\alpha_{k-1}$ is the minimal polynomial of A. Why is $\alpha_{k-1} \neq 0$?

6. Carry out the computations required by the algorithm in Problem 5 to determine the minimal polynomials of $\begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$, $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, and $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

7. Consider $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ to show that the minimal polynomials of AB and BA need not be the same. The characteristic polynomials of AB and BA are the same, however. Explain why there is this difference between the characteristic and minimal polynomials.

8. Let $A_i \in M_{n_i}$, $i = 1, 2, \ldots, k$ and let $q_i(t)$ denote the minimal polynomial of each A_i. Show that the minimal polynomial of the direct sum

$$A = \begin{bmatrix} A_1 & & & 0 \\ & A_2 & & \\ & & \ddots & \\ 0 & & & A_k \end{bmatrix}$$

is the least common multiple of $q_1(t), q_2(t), \ldots, q_k(t)$. This is the unique monic polynomial of minimum degree that is divisible by each $q_i(t)$. Notice that this argument gives a different proof for Lemma (1.3.10).

9. If $A \in M_5$ has characteristic polynomial $p_A(t) = (t-4)^3(t+6)^2$ and minimal polynomial $q_A(t) = (t-4)^2(t+6)$, what is the Jordan canonical form of A?

10. Show by direct computation that the polynomial (3.3.11) is the characteristic polynomial of the companion matrix (3.3.12). *Hint:* Use cofactors to compute the determinant.

11. Sometimes one sees the companion matrix of the polynomial (3.3.11) defined to be

$$\begin{bmatrix} -a_{n-1} & -a_{n-2} & \cdots & -a_0 \\ 1 & 0 & \cdots & 0 \\ & 1 & \ddots & \vdots \\ 0 & & \ddots\ \ 1 & 0 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 0 & 1 & & 0 \\ \vdots & & \ddots & \\ 0 & & 0 & 1 \\ -a_0 & -a_1 & \cdots & -a_{n-1} \end{bmatrix}$$

Show that both of these matrices share with (3.3.12) the property that (3.3.11) is both the minimal and characteristic polynomial of the matrix.

12. Show that there is no real 3-by-3 matrix whose minimal polynomial is x^2+1, but that there is a real 2-by-2 matrix as well as a complex 3-by-3 matrix with this property. *Hint:* Use (3.3.4).

13. Although similar matrices have the same characteristic and minimal polynomials, show that two matrices of order 4 or more can have the same minimal and characteristic polynomial without being similar. *Hint:* Consider

$$\left[\begin{array}{cc|cc} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right] \quad \text{and} \quad \left[\begin{array}{cc|cc} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right].$$

Show that 4 is the minimum order for which this can occur.

14. If $A, B \in M_n$ are similar, and if $p(t)$ is a polynomial, then $p(A)=0$ if and only if $p(B)=0$. Use the example in the preceding problem to show that it is possible to have $p(A)=0$ if and only if $p(B)=0$ for every polynomial $p(t)$ even if A and B are not similar. How can this occur?

15. Let $A \in M_n$ be a given matrix, and let $P(A)=\{p(A): p(t)$ is a polynomial$\}$. Show that $P(A)$ is a subspace of M_n and that it is even a subalgebra of M_n [$P(A)$ is closed under products]. Show that the dimension of $P(A)$ is the degree of the minimal polynomial of A.

16. If $A, B \in M_n$ have the same characteristic polynomial and the same minimal polynomial, and their minimal polynomial is the same as their characteristic polynomial, show that A and B are similar. Use this fact to show that the various alternate forms for the companion matrix noted in Problem 11 are all similar to (3.3.12).

3.4 Other canonical forms and factorizations

In addition to the Jordan canonical form, there are several other matrix factorizations that can be useful in various circumstances.

We consider first a variant of the Jordan canonical form (3.1.12) when the matrix A has only real entries. In this case, all the nonreal eigenvalues must occur in conjugate pairs. Moreover, if A is real, then $\operatorname{rank}(A-\lambda I)^k = \operatorname{rank}(\overline{A-\lambda I})^k = \operatorname{rank}(A-\bar{\lambda} I)^k$ for all $\lambda \in \mathbf{C}$ and all $k = 1, 2, \ldots$, and hence the structure of the Jordan blocks corresponding to any eigenvalue λ is the same as the structure of the Jordan blocks corresponding to the conjugate eigenvalue $\bar{\lambda}$. Thus, all the Jordan blocks of all sizes (not just the 1-by-1 blocks) corresponding to nonreal eigenvalues occur in conjugate pairs of equal size.

For example, if λ is a nonreal eigenvalue of the real matrix A, and if $J_2(\lambda)$ appears in the Jordan canonical form of A with a certain multiplicity, $J_2(\bar{\lambda})$ must also appear with the same multiplicity. The block matrix

$$\begin{bmatrix} J_2(\lambda) & 0 \\ 0 & J_2(\bar{\lambda}) \end{bmatrix} = \left[\begin{array}{cc|cc} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 0 & 0 \\ \hline 0 & 0 & \bar{\lambda} & 1 \\ 0 & 0 & 0 & \bar{\lambda} \end{array}\right] \tag{3.4.1}$$

is permutation-similar (interchange rows and columns 2 and 3) to the block matrix

$$\left[\begin{array}{cc|cc} \lambda & 0 & 1 & 0 \\ 0 & \bar{\lambda} & 0 & 1 \\ \hline 0 & 0 & \lambda & 0 \\ 0 & 0 & 0 & \bar{\lambda} \end{array}\right] = \begin{bmatrix} D(\lambda) & I \\ 0 & D(\lambda) \end{bmatrix}$$

where

$$D(\lambda) \equiv \begin{bmatrix} \lambda & 0 \\ 0 & \bar{\lambda} \end{bmatrix} \in M_2 \quad \text{and} \quad I \in M_2$$

In general, any Jordan matrix of the form

$$\begin{bmatrix} J_k(\lambda) & 0 \\ 0 & J_k(\bar{\lambda}) \end{bmatrix} \in M_{2k} \tag{3.4.2}$$

is permutation-similar to the block matrix

$$\begin{bmatrix} D(\lambda) & I & & & 0 \\ & D(\lambda) & I & & \\ & & \ddots & \ddots & \\ & & & & I \\ 0 & & & & D(\lambda) \end{bmatrix} \in M_{2k}$$

with k blocks $D(\lambda)$ on the main diagonal and $k-1$ 2-by-2 identity matrices on the superdiagonal.

Each 2-by-2 diagonal block $D(\lambda)$ is similar to a real 2-by-2 matrix

$$SD(\lambda)S^{-1} = \begin{bmatrix} a & b \\ -b & a \end{bmatrix} \equiv C(a, b) \tag{3.4.3}$$

where $\lambda = a+ib$, $a,b\in\mathbf{R}$, and $S=\begin{bmatrix} -i & -i \\ 1 & -1\end{bmatrix}$. Thus, every block pair of conjugate 2-by-2 Jordan blocks (3.4.1) with nonreal λ is similar via $\begin{bmatrix} S & 0 \\ 0 & S\end{bmatrix}$ to a real 4-by-4 block of the form

$$\left[\begin{array}{cc|cc} a & b & 1 & 0 \\ -b & a & 0 & 1 \\ \hline 0 & 0 & a & b \\ 0 & 0 & -b & a \end{array}\right] = \begin{bmatrix} C(a,b) & I \\ 0 & C(a,b)\end{bmatrix}$$

In general, every block pair of conjugate k-by-k Jordan blocks (3.4.2) with nonreal λ is similar to a real $2k$-by-$2k$ block of the form

$$C_k(a,b) \equiv \begin{bmatrix} C(a,b) & I & & \mathbf{0} \\ & C(a,b) & \ddots & \\ & & \ddots & I \\ \mathbf{0} & & & C(a,b) \end{bmatrix} \tag{3.4.4}$$

These observations lead us to the *real Jordan canonical form.*

3.4.5 **Theorem.** Each real matrix $A\in M_n(\mathbf{R})$ is similar to a block diagonal real matrix of the form

$$\begin{bmatrix} C_{n_1}(a_1,b_1) & & & & & & \mathbf{0} \\ & C_{n_2}(a_2,b_2) & & & & & \\ & & \ddots & & & & \\ & & & C_{n_p}(a_p,b_p) & & & \\ & & & & J_{n_q}(\lambda_q) & & \\ & & & & & \ddots & \\ \mathbf{0} & & & & & & J_{n_r}(\lambda_r) \end{bmatrix} \tag{3.4.6}$$

where $\lambda_k = a_k + ib_k$ is a nonreal eigenvalue of A for $k=1,2,\ldots,p$, a_k and b_k are real, and $\lambda_q,\ldots,\lambda_r$ are the real eigenvalues of A. Each real block triangular matrix $C_{n_k}(a_k,b_k)\in M_{2n_k}$ is of the form (3.4.4) and corresponds to a pair of conjugate Jordan blocks $J_{n_k}(\lambda_k), J_{n_k}(\bar{\lambda}_k)\in M_{n_k}$ with nonreal λ_k in the Jordan canonical form (3.1.12) of A. The real Jordan blocks $J_{n_k}(\lambda_k)$ in (3.4.6) are exactly the Jordan blocks in (3.1.12) with real λ_k.

We have deduced the real Jordan form of a real matrix from its general (complex) Jordan canonical form (3.1.12). This approach has the advantage of showing exactly how the sizes and numbers of the real blocks $C_{n_k}(a_k,b_k)$ are related to the complex Jordan block structure of A. The disadvantage of this approach, however, is that it is not evident that the similarity matrix that transforms A into (3.4.6) can be chosen to be real.

In fact, if A is real, there is always a *real* nonsingular matrix S such that $S^{-1}AS$ is in real Jordan form (3.4.6). One can prove this by following the three steps in the proof of the Jordan canonical form theorem in (3.1), starting with the real form (2.3.4) of Schur's triangularization theorem instead of the complex form (2.3.1). In steps 2 and 3 one can mimic the arguments in the complex case to show that one can use a real similarity in each case to reduce to modified triangular or Jordan diagonal blocks in which there may be real 2-by-2 blocks $C(a,b)$ of the form (3.4.3) on the main diagonal.

The complex Jordan canonical form (3.1.12) is a direct sum of upper triangular matrices, and the real Jordan form (3.4.6) is a direct sum of Hessenberg or "almost upper triangular" matrices since each 2-by-2 real block $C(a,b)$ has one entry below the main diagonal.

It is also possible to develop canonical forms that are a direct sum of companion matrices. These forms have some appeal for complex matrices but have the advantage that they are valid for fields other than $\mathbf{C}$, for which the Jordan canonical form is not available.

Let $A\in M_n$ be a given matrix, and let its Jordan canonical form be (3.1.12). Group together all the Jordan blocks corresponding to each distinct eigenvalue. From each group, select a Jordan block of largest order and remove it from the group. Let B_1 denote the direct sum of all these removed blocks. There will be as many direct summands in B_1 as there are distinct eigenvalues of A. Now select from the remaining blocks in each group a Jordan block of largest order and remove it from the group. Let B_2 denote the direct sum of all these blocks. There may be fewer direct summands in B_2 than in B_1 because some group of blocks may now be empty; that is, some eigenvalue of A may have only one Jordan block associated with it. Continue this process to form direct sums $B_1, B_2, B_3, \ldots, B_s$ until all the groups of Jordan blocks are empty. The sizes of B_k are monotone nonincreasing. Then $B_1\oplus B_2\oplus\cdots\oplus B_s$ is permutation-similar to the original Jordan form (3.1.12) of A.

Because of the way the direct sums B_k have been constructed, the minimal and characteristic polynomials of each B_k are the same. In fact, the characteristic (minimal) polynomial of B_1 is exactly the minimal polynomial of A. Thus, each B_k is similar to the companion matrix of its characteristic (minimal) polynomial by (3.3.15).

The characteristic (minimal) polynomials of the matrices B_k are known

as the *invariant factors* $f_k(t)$ of A. Notice that their degrees are monotone nonincreasing and that each $f_{k+1}(t)$ divides $f_k(t)$ for $k=1,2,\ldots,s-1$. The first invariant factor $f_1(t)=q_{B_1}(t)$ is the minimal polynomial of A, and the product of all the invariant factors is the characteristic polynomial of A. The invariant factors are determined in a definite way by the Jordan block structure of A, which is determined by the eigenvalues λ_i of A and the sequence of ranks of powers of $(A-\lambda_i I)$. Thus, similar matrices will have the same invariant factors and, since the invariant factors also determine the Jordan block structure of A, two matrices with the same invariant factors must be similar. Thus, the sequence of invariant factors of a matrix (which includes the minimal polynomial and determines the characteristic polynomial) is a *complete set of polynomial similarity invariants:* Two matrices $A, B \in M_n$ are similar if and only if their invariant factors are identical.

Another way to characterize the invariant factors of A is to define $f_1(t)=(t-\lambda_1)^{r_1}\cdots(t-\lambda_m)^{r_m}$ to be the minimal polynomial of A. Now delete from the Jordan form of A one Jordan block corresponding to each factor $(t-\lambda_i)^{r_i}$ of $f_1(t)$ (these are just the Jordan blocks that comprise B_1) and let $f_2(t)=(t-\lambda_1)^{s_1}\cdots(t-\lambda_m)^{s_m}$ be the minimal polynomial of the remaining Jordan form. Now delete one block corresponding to each factor $(t-\lambda_i)^{s_i}$ and let $f_3(t)$ be the minimal polynomial of what remains, and so on. The invariant factors $f_k(t)$ are just the minimal polynomials of a series of successively deflated matrices in which certain Jordan blocks are removed at each step.

The characterization of similar matrices in terms of invariant factors is conceptually pleasant, since it shows clearly why the minimal and characteristic polynomials are generally insufficient to distinguish similarity, but it really adds nothing to the criterion we already know: Two matrices are similar if and only if their Jordan canonical forms are the same.

On the other hand, this characterization leads to a new canonical form for A known as the *rational form* because the invariant factors can be computed using only rational operations on the entries of the matrix A.

3.4.7 Theorem. Every matrix $A \in M_n$ is similar to the direct sum of the companion matrices of its invariant factors.

Since we already have the Jordan canonical form, the rational form (3.4.7) for complex matrices may not seem to have any advantages. The reason for introducing it is that a version of it is true over *every* field, not just the complex numbers. Any matrix over a field $\mathbf{F}$ is similar *over* $\mathbf{F}$ to the direct sum of the companion matrices of its invariant factors, which are uniquely determined polynomials with coefficients from $\mathbf{F}$. We illustrate this for the real field.

If $A \in M_n(\mathbf{R})$ is a given real matrix, then A is similar (by a possibly complex similarity transformation) to the direct sum $B_1 \oplus B_2 \oplus \cdots \oplus B_s$, which, in turn, is similar to a direct sum of the companion matrices of its invariant factors (the characteristic polynomials of the B_k terms, $k=1,\ldots,s$). The Jordan blocks of A that correspond to nonreal eigenvalues must occur in conjugate pairs. Therefore, if a block corresponding to a nonreal eigenvalue occurs in any B_k, its conjugate must also occur in the same B_k, $k=1,\ldots,s$. Thus, each B_k has a *real* characteristic polynomial, and the form guaranteed by (3.4.7) is a real matrix. This form, the rational form for real matrices, may actually be achieved by a real similarity, for which we omit a proof.

3.4.8 Theorem. Every real matrix $A \in M_n(\mathbf{R})$ is similar over $\mathbf{R}$ to a direct sum of companion matrices of real monic polynomials $p_1(t)$, $p_2(t),\ldots,p_s(t)$ in which each $p_{k+1}(t)$ divides $p_k(t)$ for $k=1,2,\ldots,s-1$. The polynomial $p_1(t)$ is the minimal polynomial of A over $\mathbf{R}$, the product $p_1(t)\cdots p_s(t)$ is the characteristic polynomial of A, and each $p_k(t)$ is an invariant factor of A over $\mathbf{R}$. The polynomials $p_k(t)$ are uniquely determined, so two real matrices are similar over $\mathbf{R}$ if and only if they have the same invariant factors.

We emphasize that a theorem of the same form is true over the field $\mathbf{Q}$ of rational numbers or any other field. The rational form gets its name from the fact that reduction of a matrix $A \in M_n(\mathbf{F})$ to the stated form can, in principle, be accomplished by finitely many rational computations on the entries of A that stay within the field $\mathbf{F}$. Thus, if $\mathbf{F}$ is the field of rational numbers, only similarities with rational entries and polynomials with rational coefficients are used.

A different canonical form involving companion matrices can also be derived from the Jordan canonical form (3.1.12). Observe that every individual Jordan block has the property that its minimal and characteristic polynomials are identical. Thus, each Jordan block $J_{n_i}(\lambda_i)$ is similar to the companion matrix of its characteristic polynomial $(t-\lambda_i)^{n_i}$. The whole Jordan canonical form is therefore similar to a direct sum of companion matrices of these polynomials $(t-\lambda_i)^{n_i}$; these polynomials are known as the *elementary divisors* of A. Notice that there are generally more direct summands in this way of factorizing A than in the rational form; each invariant factor may yield several elementary divisors. The product of all the elementary divisors of A is the characteristic polynomial of A.

If $A \in M_n(\mathbf{R})$, compute its Jordan form and elementary divisors over $\mathbf{C}$ and notice that they must occur in conjugate pairs. If blocks $J_{n_i}(\lambda)$ and $J_{n_i}(\bar{\lambda})$ are combined as a direct sum, the resulting block has the real poly-

nomial $(t-\lambda)^{n_i}(t-\bar{\lambda})^{n_i}$ as its characteristic and minimal polynomial, and hence it is similar to the real companion matrix of $[t^2-(2\operatorname{Re}\lambda)t+|\lambda|^2]^{n_i}$. The latter polynomial is a real elementary divisor of A. Powers of real linear factors also occur as elementary divisors for each real eigenvalue of A.

The canonical form associated with elementary divisors is, with inevitable confusion, usually called the *rational canonical form*.

3.4.9 **Theorem.** Every matrix $A \in M_n(\mathbf{R})$ is similar over $\mathbf{R}$ to the direct sum of companion matrices of its (real) elementary divisors.

The same sort of result is true over any field $\mathbf{F}$: Every $A \in M_n(\mathbf{F})$ is similar over $\mathbf{F}$ to the direct sum of companion matrices of its elementary divisors, which are polynomials with coefficients from $\mathbf{F}$.

As an example, consider the matrices

$$A_1=[1], \qquad A_2=\begin{bmatrix} 0 & -4 \\ 1 & 4 \end{bmatrix}, \qquad A_3=\begin{bmatrix} 0 & -9 \\ 1 & 0 \end{bmatrix}, \qquad A_4=\begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$$

and let $A=A_1\oplus A_2\oplus A_3\oplus A_3\oplus A_4 \in M_9$. Then the rational canonical form of A over $\mathbf{R}$ is $A=A_1\oplus A_2\oplus A_3\oplus A_3\oplus[4]\oplus[4]$, and the elementary divisors are $x-1$, $(x-2)^2$, x^2+9, x^2+9, $x-4$, $x-4$. Over $\mathbf{C}$, the rational canonical form of A is $A_1\oplus A_2\oplus[3i]\oplus[3i]\oplus[-3i]\oplus[-3i]\oplus[4]\oplus[4]$, and the elementary divisors are $x-1$, $(x-2)^2$, $x-3i$, $x-3i$, $x+3i$, $x+3i$, $x-4$, $x-4$. The rational form of A over $\mathbf{R}$ is (the companion matrix of $A_1\oplus A_2\oplus A_3\oplus[4]$)$\oplus$(the companion matrix of $A_3\oplus[4]$) and the invariant factors are

$$f_1(t)=(t-1)(t-2)^2(t^2+9)(t-4) \quad \text{and} \quad f_2(t)=(t^2+9)(t-4)$$

Over $\mathbf{C}$, the rational form of A is (the companion matrix of $A_1\oplus A_2\oplus[3i]\oplus[-3i]\oplus[4]$)$\oplus$(the companion matrix of $[3i]\oplus[-3i]\oplus[4]$). Notice that the two direct summands and the invariant factors are the same whether A is thought of as a matrix in $M_n(\mathbf{R})$ or $M_n(\mathbf{C})$. This is not true of the rational canonical form and the elementary divisors. See Problems 2 and 3 at the end of this section.

We do not use the real Jordan form, the rational form, the rational canonical form, invariant factors, or elementary divisors in any essential way in the rest of the book. We have discussed them here only because of their historical importance and their necessity when one does matrix analysis over fields other than $\mathbf{C}$.

There are many other useful canonical forms and matrix factorizations:

(a) The polar decomposition: Every $A \in M_n$ can be written as $A=PU$, where $P \in M_n$ is a positive semidefinite matrix whose rank is the same as that of A, and $U \in M_n$ is a unitary matrix. See (7.3.3). Every nonsingular matrix $A \in M_n$ can also be written as $A=GQ$,

where $G \in M_n$ is a (complex) symmetric matrix ($G=G^T$) and $Q \in M_n$ is a (complex) orthogonal matrix ($QQ^T=I$).

(b) The singular value decomposition: Every $A \in M_n$ can be written as $A=V\Sigma W^*$, where $V, W \in M_n$ are unitary, and $\Sigma \in M_n$ is a diagonal matrix with nonnegative main diagonal entries and the rank of Σ is the same as the rank of A. See (7.3.5).

(c) The triangular factorization: Every $A \in M_n$ can be written as $A=URU^*$, where $U \in M_n$ is unitary and $R \in M_n$ is upper triangular. Every real matrix $A \in M_n(\mathbf{R})$ can be written as $A=QRQ^T$, where $Q \in M_n(\mathbf{R})$ is orthogonal and $R \in M_n(\mathbf{R})$ is an upper Hessenberg matrix with a special structure. See (2.3.5).

(d) Every Hermitian matrix $A \in M_n$ can be written as $A=SI(A)S^*$, where $S \in M_n$ is nonsingular and $I(A) \in M_n$ is a diagonal matrix with $+1$, -1, or 0 main diagonal entries. The number of $+1$ (-1) entries in $I(A)$ is the same as the number of positive (negative) eigenvalues of A; the number of 0 entries is equal to $n-\operatorname{rank} A$. See (4.5.8).

(e) Every normal matrix $A \in M_n$ can be written as $A=U\Lambda U^*$, where $U \in M_n$ is unitary and $\Lambda \in M_n$ is a diagonal matrix whose main diagonal entries are the eigenvalues of A. Every real normal matrix $A \in M_n(\mathbf{R})$ can be written as $A=QDQ^T$, where $Q \in M_n(\mathbf{R})$ is orthogonal and $D \in M_n(\mathbf{R})$ is a block diagonal matrix with a special structure. See (2.5.8).

(f) Every matrix $A \in M_n$ such that $A=A^T$ can be written as $A=SK(A)S^T$, where $S \in M_n$ is nonsingular and $K(A) \in M_n$ is a diagonal matrix whose main diagonal entries are $+1$ or 0 and whose rank is equal to the rank of A. See (4.5.12).

(g) Every matrix $A \in M_n$ such that $A=A^T$ can be written as $A=U\Sigma U^T$, where $U \in M_n$ is unitary and Σ is a diagonal matrix with nonnegative main diagonal entries. The rank of Σ is equal to the rank of A. See (4.4.4).

(h) Every unitary matrix $U \in M_n$ can be written as $U=Qe^{iE}$ and every (complex) orthogonal matrix $P \in M_n$ can be written as $P=Qe^{iF}$ where $Q, E, F \in M_n(\mathbf{R})$, Q is real orthogonal ($QQ^T=I$), E is real symmetric ($E=E^T$), and F is real skew-symmetric ($F=-F^T$).

(i) Every matrix $A \in M_n$ can be written as $A=SU\Sigma U^TS^{-1}$, where S is nonsingular, U is unitary, and Σ is a diagonal matrix with nonnegative main diagonal entries. See (4.4.10).

Problems

1. Compute the minimal and characteristic polynomials, the invariant factors, elementary divisors, rational form, and rational canonical form over $\mathbf{R}$ and $\mathbf{C}$ for

$$\begin{bmatrix} 0 & 0 & -1 \\ -1 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

2. Let $A \in M_n(\mathbf{R})$. Suppose $q(t)$ is the minimal polynomial of A over $\mathbf{R}$ and $f(t)$ is the minimal polynomial of A over $\mathbf{C}$. Why is degree $f(t) \leq$ degree $q(t)$? Why must $f(t)$ divide $q(t)$? Show that $f(t) = q(t)$ by considering $f(t) = p_1(t) + ip_2(t)$, where $p_1(t)$ and $p_2(t)$ have real coefficients. Why is $p_1(A) = p_2(A) = 0$?

3. Use Theorem (3.4.7) to show that if $A, B \in M_n(\mathbf{F})$ and if $\mathbf{F}$ is a subfield of $\mathbf{C}$ ($\mathbf{F} = \mathbf{R}$ or $\mathbf{Q}$, for example), then A and B are similar over $\mathbf{F}$ if and only if they are similar over $\mathbf{C}$. *Hint:* Show that the rational form of A over $\mathbf{F}$ is the same as the rational form of A over $\mathbf{C}$, and similarly for B. How does this result generalize Problem 2?

4. Let $A \in M_n(\mathbf{R})$, and suppose $A^2 = -I$. Show that n must be even, and that there is a real nonsingular matrix $S \in M_n$ such that

$$S^{-1}AS = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}$$

where each identity $I \in M_{n/2}$.

Further Readings. The rational forms mentioned in this section are quite classical and can be found in more detail in [HKu]. The real Jordan canonical form seems also to have been known to matrix theorists for some time, but presentations of it are not common. A statement of the real Jordan form can be found in [Kow], for example. See [New] for a discussion of canonical forms of matrices with rational or integer entries. See [Gan], vol. 2; [Gant]; and [HJ] for additional details on special canonical forms.

3.5 Triangular factorizations

If a linear system $Ax = b$ has a nonsingular triangular (0.9.3) coefficient matrix $A \in M_n$, computation of the unique solution x is remarkably easy. If, for example, A is upper triangular,

$$A = \begin{bmatrix} a_{11} & \cdots\cdots & a_{1n} \\ & a_{22} & \vdots \\ & & \ddots & \vdots \\ 0 & & & a_{nn} \end{bmatrix}$$

then $\det A = a_{11} a_{22} \cdots a_{nn} \neq 0$ and *back substitution* is used: $a_{n,n}x_n = b_n$ determines x_n; $a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n = b_{n-1}$ is then one equation in one unknown, which determines x_{n-1}; and, in general, each of the sequence of equations

$$\sum_{j=i}^{n} a_{ij}x_j = b_i, \qquad i = n, n-1, n-2, \ldots, 2, 1$$

is one equation in one unknown (once $x_{i+1}, \ldots, x_n$ have been determined), which determines x_i.

Exercise. Count the number of scalar multiplication and division operations necessary to solve $Ax = b$, if $A \in M_n$ is nonsingular and upper triangular, if one uses back substitution.

Exercise. Describe *forward substitution* as a solution technique for $Ax = b$ if $A \in M_n$ is nonsingular and lower triangular.

If $A \in M_n$ is nonsingular but not triangular, notice that it is almost as convenient to solve $Ax = b$ if A is given in factored form as

$$A = LU$$

in which L is lower triangular and U is upper triangular.

Exercise. If $A = LU$, as above, is nonsingular, show that L and U must both be nonsingular and must therefore have nonzero diagonal entries.

In order to solve $Ax = b$, we may first solve

$Ly = b$ by forward substitution

and then

$Ux = y$ by backward substitution

and the necessary computational effort is only twice as great as in the simple triangular case. Thus, factorizations such as LU can be helpful in solving linear systems if the cost of achieving them is not too great. They are also appropriate to mention here, as they are special forms into which a matrix can be put – motivated now, not by eigenvalues, but by linear systems.

3.5.1 **Lemma.** Suppose that $A \in M_n$ can be written

$$A = LU$$

with $L \in M_n$ lower triangular and $U \in M_n$ upper triangular. For any partition

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \qquad L = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix}, \qquad U = \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix}$$

with $A_{11}, L_{11}, U_{11} \in M_k$, $k \leq n$, we have

$$L_{11} U_{11} = A_{11}$$

$$L_{11} U_{12} = A_{12}, \qquad L_{21} U_{11} = A_{21}$$

and

$$L_{21} U_{12} + L_{22} U_{22} = A_{22}$$

In particular, the upper left blocks of L and U must form a factorization, of the same type, of the corresponding block of A.

Exercise. Verify (3.5.1) by carrying out the partitioned multiplication.

3.5.2 **Theorem.** Suppose that $A \in M_n$ and that rank $A = k$. If

$$\det A(\{1, \ldots, j\}) \neq 0, \qquad j = 1, \ldots, k$$

then A may be factored as

$$A = LU$$

with $L \in M_n$ lower triangular and $U \in M_n$ upper triangular. Furthermore, the factorization may be chosen so that either L or U is nonsingular; both L and U may be chosen nonsingular if and only if $k = n$, that is, if and only if A is nonsingular.

Proof: We first show that, under the assumption on leading minors, $A(\{1, \ldots, k\})$ may be factored as $L(\{1, \ldots, k\})U(\{1, \ldots, k\})$, with both nonsingular. It is possible to solve for the relevant entries of L and U, one by one. Let $L = [l_{ij}]$ and $U = [u_{ij}]$. Set $u_{11} = 1$, and let $l_{i1} = a_{i1}$, $i = 1, \ldots, k$. Solve for

$$u_{1j} = \frac{a_{1j}}{l_{11}}, \qquad j = 2, \ldots, k$$

Continue. Set $u_{22} = 1$ and let $l_{i2} = a_{i2} - l_{i1} u_{12}$, $i = 2, \ldots, k$. Solve for

$$u_{2j} = \frac{a_{2j} - l_{21} u_{1j}}{l_{22}}, \qquad j = 3, \ldots, k$$

Continue, letting successive diagonal entries of U be 1 and then solving for the next column of $L(\{1, \ldots, k\})$ and then the next row of $U(\{1, \ldots, k\})$.

Each time there is one equation in one unknown to be solved. This equation will be solvable since each l_{ii} is nonzero [because $\det L(\{1, \ldots, i\}) \times \det U(\{1, \ldots, i\}) = \det A(\{1, \ldots, i\})$ by (3.5.1)]. This completes the factorization of $A(\{1, \ldots, k\})$.

Partition A as in (3.5.1). Since rank $A = k =$ rank A_{11}, we see that the rows of $[A_{21}\ A_{22}]$ are unique linear combinations of the rows of $[A_{11}\ A_{12}]$, that is,

$$A_{21} = BA_{11} \quad \text{and} \quad A_{22} = BA_{12}$$

for some uniquely determined $B \in M_{n-k,k}$. Now partition the desired L and U, also as in (3.5.1), noting that nonsingular L_{11} and U_{11} have been determined. We may then use (3.5.1) to solve for

$$U_{12} = L_{11}^{-1} A_{12} \quad \text{and} \quad L_{21} = A_{21} U_{11}^{-1}$$

Then

$$A_{22} = L_{21} U_{12} + L_{22} U_{22} = A_{21} U_{11}^{-1} L_{11}^{-1} A_{12} + L_{22} U_{22} = BA_{11} A_{11}^{-1} A_{12} + L_{22} U_{22}$$
$$= A_{22} + L_{22} U_{22}$$

To complete the factorization, it is necessary and sufficient that

$$L_{22} U_{22} = 0$$

We may, for example, choose L_{22} (respectively U_{22}) to be any nonsingular lower (respectively upper) triangular matrix in M_{n-k} we like and choose U_{22} (respectively L_{22}) to be 0. Since L_{11} and U_{11} are nonsingular, either L or U may be chosen to be nonsingular. If $k = n$, $L = L_{11}$ and $U = U_{11}$ will be nonsingular; if $k < n$, not both L and U can be nonsingular because A is singular. This completes the proof. □

3.5.3 **Example.** Not every matrix has an LU factorization. Consider

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

If A could be written as

$$A = LU = \begin{bmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix}$$

$l_{11} u_{11} = 0$ would require that one of L or U be singular, but $LU = A$ is nonsingular.

Exercise. Show that a nonsingular matrix that has an upper left k-by-k singular principal submatrix cannot have an LU factorization.

3.5.4 **Example.** It is possible for $A \in M_n$ to have an LU-factorization without satisfying the principal minor conditions of (3.5.2). For example,

$$\begin{bmatrix} 0 & 0 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

has rank 1, but its 1, 1 entry is 0.

Exercise. The LU factorization in (3.5.4) is not unique, even if the diagonal entries of U are required to be 1. Give several other factorizations of $\begin{bmatrix} 0 & 0 \\ 1 & 2 \end{bmatrix}$.

It should now be clear that an LU factorization of a given matrix can be highly nonunique and may or may not exist. Much of the trouble, however, arises from singularity, either in A or in its leading principal submatrices. Using the tools of (3.5.1) and (3.5.2), however, we can give a full description in the nonsingular case, and we can impose a normalization that makes the factorization unique (canonical).

3.5.5 **Corollary.** Suppose that $A \in M_n$ is nonsingular. Then A may be written as

$$A = LU$$

with $L \in M_n$ lower triangular and $U \in M_n$ upper triangular, if and only if

$$\det A(\{1, \ldots, j\}) \neq 0, \qquad j = 1, \ldots, n$$

Furthermore, L and U are nonsingular and the factorization is essentially unique. The matrix A may be written as

$$A = L'DU'$$

in which L' (respectively U') $\in M_n$ is lower (respectively upper) triangular with all diagonal entries equal to 1, and D is a nonsingular diagonal matrix determined by

$$\det D(\{1, \ldots, j\}) = \det A(\{1, \ldots, j\}), \qquad j = 1, \ldots, n$$

The factors L', U', and D are uniquely determined by A.

Exercise. Provide details of a proof of (3.5.5) using (3.5.1), (3.5.2), and prior exercises.

Returning to the solution of the linear system

$$Ax = b$$

suppose that $A \in M_n$ cannot be factored as LU, but can be factored as PLU, in which $P \in M_n$ is a *permutation matrix* (0.9.5) and L and U are lower and upper triangular, as before. This amounts to a reordering of the equations prior to factorization. In this event, solution of $Ax = b$ is still quite simple via

$$Ly = P^T b \quad \text{and} \quad Ux = y$$

It is worth realizing that any nonsingular $A \in M_n$ may be so factored, and *any* $A \in M_n$ may be factored as $PLUQ$, in which $Q \in M_n$ is also a permutation.

3.5.6 **Lemma.** Let $A \in M_k$ be nonsingular. Then there is a permutation matrix $P \in M_k$ such that

$$\det(P^T A)(\{1, \ldots, j\}) \neq 0, \qquad j = 1, \ldots, k$$

Note that $P^T A$ is just a reordering of the rows of A.

Proof: The demonstration is by induction on k. If $k = 1$ or 2, the result is clear by inspection; suppose that it is valid up to and including $k-1$. Consider a nonsingular $A \in M_k$ and delete its last column. The remaining $k-1$ columns are linearly independent and hence contain $k-1$ linearly independent rows. Permute these rows to the first $k-1$ positions and apply the induction hypothesis to the nonsingular upper $(k-1)$-by-$(k-1)$ submatrix. This determines a desired overall permutation. Since $P^T A$ is nonsingular, the proof is complete. □

3.5.7 **Theorem.** Let $A \in M_n$. There exist permutation matrices $P, Q \in M_n$, a lower triangular matrix $L \in M_n$, and an upper triangular matrix $U \in M_n$ such that

$$A = PLUQ$$

If A is nonsingular, one may take $Q = I$ and A may be written as

$$A = PLU$$

Proof: If rank $A = k$, A has a k-by-k nonsingular submatrix (0.4.4d), which may, by permutation of rows and columns, be permuted into the upper left corner. Now apply (3.5.6) to the upper left corner and apply (3.5.2) to achieve a factorization of the first type. If A is nonsingular, (3.5.6) indicates that permutation on the right is unnecessary in order to apply (3.5.2), which verifies the second factorization and completes the proof. □

Problems

1. The theory developed in this section is in terms of LU, with L lower triangular and U upper triangular. Show that a parallel theory of UL factorization may be developed, but that the factors will, in general, be different.

2. Recall from Problem 3 of Section (2.6) that the QR factorization (2.6.1) of an arbitrary $A \in M_n$ may be achieved efficiently by $n-1$ Householder transformations. Here, Q is unitary and R is upper triangular. Describe how $Ax = b$ may be solved if A is factored in QR form.

3. Show that $A \in M_n$ may be written as

$$A = LP_0U$$

in which $L \in M_n$ is a nonsingular lower triangular matrix, $U \in M_n$ is a nonsingular upper triangular matrix, and P_0 is a sub-permutation matrix [a permutation matrix with as many of the 1's replaced by 0's as the rank of A is less than n]. *Hint:* Use elementary row and column operations.

4. If the leading principal minors of $A \in M_n$ are all nonzero, describe how A may be LU-factored using type 3 elementary row operations to zero out entries below the diagonal.

5. (Lanczos tridiagonalization algorithm.) Let $A \in M_n$ and $x \in \mathbf{C}^n$ be given. Define $X = [x\ Ax\ A^2x \ldots A^{n-1}x]$. The columns of X are said to form a *Krylov sequence.* Assume that X is nonsingular. (a) Show that $X^{-1}AX$ is a companion matrix (3.3.12) for the characteristic polynomial of A. (b) If $R \in M_n$ is any given nonsingular upper triangular matrix and $S \equiv XR$, show that $S^{-1}AS$ is in upper Hessenberg form. (c) Let $y \in \mathbf{C}^n$ and define $Y = [y\ A^*y\ (A^*)^2y \ldots (A^*)^{n-1}y]$. Suppose that Y is nonsingular and that Y^*X can be written as LDU, in which L is lower triangular and U is upper triangular and nonsingular, and D is diagonal and nonsingular. Show that there exist nonsingular upper triangular matrices R and T such that $(XR)^{-1} = T^*Y^*$ and such that T^*Y^*AXR is tridiagonal and similar to A. (d) If $A \in M_n$ is Hermitian, use the above ideas to specify an algorithm to produce a tridiagonal Hermitian matrix that is similar to A.

6. Two matrices $A, B \in M_{m,n}$ are said to be *equivalent* if there are nonsingular matrices $S \in M_m$ and $T \in M_n$ such that

$$B = SAT$$

(a) Show that this notion of equivalence is an equivalence relation on $M_{m,n}$. (b) Show that every matrix $A \in M_{m,n}$ is equivalent to a matrix of the form $\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \in M_{m,n}$, $I \in M_k$, $k \leq \min\{m, n\}$. *Hint:* use elementary row

operations to attain row-reduced echelon form and then use elementary column operations on the result. (c) Show that two matrices in $M_{m,n}$ are equivalent if and only if they have the same rank. (d) Suppose that $A \in M_{m,n}$ is equivalent to the special form indicated in (b), $S\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}T = A$. Develop the solution theory for the linear system $Ax = b$ in terms of equivalence.

Further Reading. Problem 5 above is adapted from [Ste], where additional information about the numerical application of LU factorizations may be found.

CHAPTER 4

Hermitian and symmetric matrices

4.0 Introduction

4.0.1 **Example.** If $f: D \to \mathbf{R}$ is a twice continuously differentiable function on some domain $D \subset \mathbf{R}^n$, the real matrix

$$H(x) = [h_{ij}(x)] \equiv \left[\frac{\partial^2 f(x)}{\partial x_i \, \partial x_j}\right] \in M_n$$

is known as the *Hessian* of f. It is a function of x and plays an important role in the theory of optimization because it can be used to determine if a critical point is a relative maximum or minimum [see (7.0)].

For our purposes here, the only property of $H = H(x)$ that interests us follows from the important fact that the mixed partials are equal; that is,

$$\frac{\partial^2 f}{\partial x_i \, \partial x_j} = \frac{\partial^2 f}{\partial x_j \, \partial x_i} \quad \text{for all} \quad i, j = 1, 2, \ldots, n$$

In terms of the Hessian matrix $H = [h_{ij}]$, this means that $h_{ij} = h_{ji}$ for all $i, j = 1, 2, \ldots, n$; that is, $H = H^T$. A matrix $A \in M_n$ such that $A = A^T$ is said to be *symmetric*. Thus, the Hessian matrix of a real-valued twice continuously differentiable function is always a real symmetric matrix.

4.0.2 **Example.** As a second example, let $A = [a_{ij}] \in M_n$ be a given matrix with real or complex entries, and consider the quadratic form on $\mathbf{R}^n$ or $\mathbf{C}^n$ generated by A:

$$Q(x) \equiv x^T A x = \sum_{i,j=1}^{n} a_{ij} x_i x_j$$
$$= \sum_{i,j=1}^{n} \tfrac{1}{2}(a_{ij} + a_{ji}) x_i x_j$$
$$= x^T [\tfrac{1}{2}(A + A^T)] x$$

Thus, A and $\frac{1}{2}(A+A^T)$ both generate the same quadratic form, and the latter matrix is symmetric. To study real or complex quadratic forms, therefore, it suffices to study only those forms generated by symmetric matrices. Real quadratic forms arise naturally in physics, for example, as an expression for the inertia of a physical body.

4.0.3 **Example.** As a third example, consider a second-order linear partial differential operator L defined by

$$Lf(x) \equiv \sum_{i,j=1}^{n} a_{ij}(x) \frac{\partial^2 f(x)}{\partial x_i \, \partial x_j} \tag{4.0.4}$$

The coefficients $a_{ij}(x)$ and the function $f(x)$ are assumed to be defined on the same domain $D \subset \mathbf{R}^n$, and f should be twice continuously differentiable on D. The operator L is associated in a natural way with a matrix. The matrix $A = [a_{ij}(x)]$ need not be symmetric, but since the mixed partial derivatives of f are equal, we have

$$Lf = \sum_{i,j=1}^{n} a_{ij}(x) \frac{\partial^2 f}{\partial x_i \, \partial x_j} = \sum_{i,j=1}^{n} \frac{1}{2}\left[a_{ij}(x) \frac{\partial^2 f}{\partial x_i \, \partial x_j} + a_{ji}(x) \frac{\partial^2 f}{\partial x_j \, \partial x_i}\right]$$
$$= \sum_{i,j=1}^{n} \frac{1}{2}[a_{ij}(x) + a_{ji}(x)] \frac{\partial^2 f}{\partial x_i \, \partial x_j}$$

Thus, the symmetric matrix $\frac{1}{2}(A+A^T)$ yields the same operator L as the matrix A, and for the study of real or complex linear partial differential operators of the form (4.0.4) it suffices to consider only symmetric coefficient matrices.

4.0.5 **Example.** Consider an undirected graph Γ; that is, Γ consists of a collection N of "nodes" $\{P_1, P_2, \ldots, P_n\}$ and a collection E of unordered pairs of nodes called "edges," $E = \{\{P_{i_1}, P_{j_1}\}, \{P_{i_2}, P_{j_2}\}, \ldots\}$. The graph Γ can be described very succinctly by its so-called *adjacency matrix* $A = [a_{ij}]$. Here,

$$a_{ij} = \begin{cases} 1 & \text{if } \{P_i, P_j\} \in E \\ 0 & \text{otherwise} \end{cases}$$

Since Γ is undirected, A is a real symmetric matrix; that is, $A^T = A$.

4.0.6 **Example.** Let $A = [a_{ij}] \in M_n$ be a real matrix and consider the real bilinear form

$$Q(x, y) = y^T A x = \sum_{i,j=1}^{n} a_{ij} y_i x_j, \qquad x, y \in \mathbf{R}^n \tag{4.0.7}$$

which reduces to the ordinary inner product when $A = I$. If we want to have $Q(x, y) = Q(y, x)$ for all x, y, then it is necessary and sufficient that $a_{ij} = a_{ji}$ for all $i, j = 1, \ldots, n$. To show this, it suffices to observe that if $x = e_j$ and $y = e_i$, then $Q(e_j, e_i) = a_{ij}$ and $Q(e_i, e_j) = a_{ji}$. Thus, symmetric real bilinear forms are naturally associated with symmetric real matrices.

Now let $A = [a_{ij}] \in M_n$ be a real or complex matrix, and consider the complex form

$$H(x, y) = y^* A x = \sum_{i,j=1}^{n} a_{ij} \bar{y}_i x_j, \qquad x, y \in \mathbf{C}^n \tag{4.0.8}$$

which, like (4.0.7), reduces to the ordinary inner product when $A = I$. This form is no longer bilinear but is linear in the first variable and "conjugate linear" in the second variable ($H(ax, by) = a\bar{b}H(x, y)$) just like the complex Euclidean inner product. Such forms are sometimes called *sesquilinear*. If we want to have $H(x, y) = \overline{H(y, x)}$ like the inner product, then the same argument as in the previous case shows that it is necessary and sufficient to have $a_{ij} = \bar{a}_{ji}$; that is, $A = \bar{A}^T \equiv A^*$. Notice that if A is real, then $A^* = A^T$.

The class of matrices $A \in M_n$ such that $A = A^*$ is in many respects the natural generalization to $M_n(\mathbf{C})$ of the class of real symmetric matrices. Such matrices are called *Hermitian*; notice that a real Hermitian matrix is a real symmetric matrix. The class of complex nonreal symmetric matrices fails to have many important properties of the class of real symmetric matrices. In this chapter we shall study complex Hermitian and symmetric matrices and will indicate by specialization what happens in the real symmetric case.

4.1 Definitions, properties, and characterizations of Hermitian matrices

4.1.1 **Definition.** A matrix $A = [a_{ij}] \in M_n$ is said to be *Hermitian* if $A = A^*$, where $A^* \equiv \bar{A}^T = [\bar{a}_{ji}]$. It is *skew-Hermitian* if $A = -A^*$.

Some observations for $A, B \in M_n$:

1. $A + A^*$, AA^*, and A^*A are all Hermitian for all $A \in M_n$.
2. If A is Hermitian, then A^k is Hermitian for all $k = 1, 2, 3, \ldots$. If A is nonsingular as well, then A^{-1} is Hermitian.

3. If A, B are Hermitian, then $aA+bB$ is Hermitian for all real scalars a, b.
4. $A-A^*$ is skew-Hermitian for all $A \in M_n$.
5. If A, B are skew-Hermitian, then $aA+bB$ is skew-Hermitian for all real scalars a, b.
6. If A is Hermitian, then iA is skew-Hermitian.
7. If A is skew-Hermitian, then iA is Hermitian.
8. Any $A \in M_n$ can be written as
$$A = \tfrac{1}{2}(A+A^*) + \tfrac{1}{2}(A-A^*) \equiv H(A)+S(A)$$
where $H(A) = \frac{1}{2}(A+A^*)$ is the *Hermitian part* of A, and $S(A) = \frac{1}{2}(A-A^*)$ is the *skew-Hermitian part* of A.
9. If A is Hermitian, the main diagonal entries of A are all real. In order to specify the n^2 elements of A one may specify freely any n real numbers (for the main diagonal entries) and any $\frac{1}{2}n(n-1)$ complex numbers (for the off-diagonal entries).

4.1.2 **Theorem.** Each $A \in M_n$ can be written uniquely as $A = S+iT$, where both S and T are Hermitian. It can also be written uniquely as $A = B+C$, where B is Hermitian and C is skew-Hermitian.

Proof: Write $A = \frac{1}{2}(A+A^*) + i[(-i/2)(A-A^*)]$ and observe that both $S = \frac{1}{2}(A+A^*)$ and $T = (-i/2)(A-A^*)$ are Hermitian. For the uniqueness assertion, observe that if $A = E+iF$ with both E and F Hermitian, then
$$2S = A+A^* = (E+iF)+(E+iF)^* = E+iF+E^*-iF^* = 2E$$
so $E = S$. Similarly, one shows that $F = T$. The assertions about the representation $A = B+C$ are proved in the same way. □

The foregoing observations suggest that if one thinks of M_n being analogous to the complex numbers, then the Hermitian matrices are analogous to the real numbers. The analog of the operation of complex conjugation in $\mathbf{C}$ is the * operation (adjoint) on M_n. A real number is a complex number z such that $z = \bar{z}$; a Hermitian matrix is a matrix $A \in M_n$ such that $A = A^*$. Just as every complex number z can be written as $z = s+it$ with $s, t \in \mathbf{R}$, every complex matrix A can be written uniquely as $A = S+iT$ with S and T Hermitian. There are some further properties that strengthen this analogy.

4.1.3 **Theorem.** Let $A \in M_n$ be Hermitian. Then

(a) x^*Ax is real for all $x \in \mathbf{C}^n$;
(b) All the eigenvalues of A are real; and
(c) S^*AS is Hermitian for all $S \in M_n$.

Proof: One computes $(\overline{x^*Ax}) = (x^*Ax)^* = x^*A^*x = x^*Ax$, so x^*Ax equals its complex conjugate and hence is real. If $Ax = \lambda x$ and $x^*x = 1$, then $\lambda = \lambda x^*x = x^*\lambda x = x^*Ax$ is real by (a). Finally, $(S^*AS)^* = S^*A^*S = S^*AS$, so S^*AS is always Hermitian.

Exercise. What does each of the foregoing properties of a Hermitian matrix $A \in M_n$ mean when $n = 1$?

Each of the properties in (4.1.3) is actually (almost) a characterization of Hermitian matrices.

4.1.4 **Theorem.** Let $A = [a_{ij}] \in M_n$ be given. Then A is Hermitian if and only if at least one of the following holds:

(a) x^*Ax is real for all $x \in \mathbf{C}^n$;
(b) A is normal and all the eigenvalues of A are real; or
(c) S^*AS is Hermitian for all $S \in M_n$.

Proof: It suffices to prove only the sufficiency of each condition. If x^*Ax is real for all $x \in \mathbf{C}^n$, then $(x+y)^*A(x+y) = (x^*Ax+y^*Ay)+(x^*Ay+y^*Ax)$ is real for all $x, y \in \mathbf{C}^n$. Since x^*Ax+y^*Ay is real by assumption, we conclude that x^*Ay+y^*Ax is real for all $x, y \in \mathbf{C}^n$. If we choose $x = e_k$ and $y = e_j$, this says that $a_{kj}+a_{jk}$ is real, so $\operatorname{Im} a_{kj} = -\operatorname{Im} a_{jk}$. If we choose $x = ie_k$ and $y = e_j$, this says that $-ia_{kj}+ia_{jk}$ is real, so $\operatorname{Re} a_{kj} = \operatorname{Re} a_{jk}$. These two identities together are equivalent to having $a_{kj} = \bar{a}_{jk}$, and since j, k are arbitrary we conclude that $A = A^*$.

If A is normal, it is unitarily diagonalizable, so $A = U\Lambda U^*$ with $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, a diagonal matrix formed from the eigenvalues of A. In general, we have $A^* = U\bar{\Lambda}U^*$, but if Λ is real, we have $A^* = U\Lambda U^* = A$. The last condition implies that A is Hermitian by choosing $S = I$. □

Since a Hermitian matrix is obviously normal ($AA^* = A^2 = A^*A$), all the results about normal matrices in Chapter 2 apply. For example, eigenvectors corresponding to distinct eigenvalues are orthogonal; there is a complete set of orthonormal eigenvectors; Hermitian matrices are unitarily diagonalizable; and so forth.

For reference, we formally state the following important result.

4.1.5 **Theorem** (the spectral theorem for Hermitian matrices). Let $A \in M_n$ be given. Then A is Hermitian if and only if there is a unitary matrix $U \in M_n$ and a real diagonal matrix $\Lambda \in M_n$ such that $A = U\Lambda U^*$. Moreover, A is real and Hermitian (i.e., real symmetric) if and only if there is a real orthogonal matrix $P \in M_n$ and a real diagonal matrix $\Lambda \in M_n$ such that $A = P\Lambda P^T$.

Although a real linear combination of Hermitian matrices is always Hermitian, a complex linear combination need not be. For example, if A is Hermitian, iA is Hermitian only if $A=0$. Furthermore, if A and B are Hermitian, then $(AB)^*=B^*A^*=BA$, so AB is Hermitian if and only if A and B commute.

One of the most famous results about commuting Hermitian matrices (because of an important generalization to operators in quantum mechanics) is the following special case of Theorem (2.5.5).

4.1.6 **Theorem.** Let $\mathfrak{F}$ be a given family of Hermitian matrices. There exists a unitary matrix U such that UAU^* is diagonal for all $A\in\mathfrak{F}$ if and only if $AB=BA$ for all $A,B\in\mathfrak{F}$.

A Hermitian matrix A has the property that A is *equal* to A^*. One way to generalize the notion of a Hermitian matrix is to consider the class of matrices such that A is *similar* to A^*. The following theorem characterizes this class in several ways, the first of which says that such matrices must be similar, but not necessarily unitarily similar, to a real, but not necessarily diagonal, matrix.

4.1.7 **Theorem.** Let $A\in M_n$ be given. The following statements are equivalent:

(a) A is similar to a matrix $B\in M_n(\mathbf{R})$;
(b) A is similar to A^*;
(c) A is similar to A^* via a Hermitian similarity transformation;
(d) $A=HK$, in which $H,K\in M_n$ are Hermitian with at least one nonsingular; and
(e) $A=HK$, in which $H,K\in M_n$ are Hermitian.

Proof: First note that (a) and (b) are equivalent: if (a) holds, then $S^{-1}AS=B=T^{-1}B^TT=T^{-1}B^*T=T^{-1}S^*A^*(S^{-1})^*T$, which means that $A^*=(ST^{-1}S^*)^{-1}A(ST^{-1}S^*)$ or that (b) holds. If (b) holds, then A and A^* have the same Jordan canonical form. Since A and A^T are similar for any matrix, this means that if J is the Jordan matrix of A, then J must be similar to $\bar{J}$. Consequently, for each Jordan block $J_k(\lambda)$ in J there is a corresponding Jordan block $J_k(\bar{\lambda})$ (of the same size) in $\bar{J}$. If λ is real, this gives no information, but if λ is not real, it means that the Jordan blocks of A corresponding to each nonreal eigenvalue and its conjugate must occur in matching pairs. Using the argument leading to (3.4.5), we conclude that J must be similar to a direct sum of real matrices of the form (3.4.4), and hence (a) follows.

To verify that (b) implies (c), suppose that $S^{-1}AS=A^*$ and observe that $T^{-1}AT=A^*$ if $T=\alpha S$ for any nonzero $\alpha=re^{i\theta}\in\mathbf{C}$. Thus, $AT=TA^*$ or, equivalently, $AT^*=T^*A^*$. Adding these two identities produces the identity $A(T+T^*)=(T+T^*)A^*$, and if $T+T^*$ were nonsingular, this would mean that A is similar to A^* via the Hermitian matrix $T+T^*$. But α may be chosen so as to make $T+T^*$ nonsingular, since $T+T^*$ is nonsingular if and only if $T^{-1}(T+T^*)=I+T^{-1}T^*$ is, if and only if $-1\notin\sigma(T^{-1}T^*)$. However, $T^{-1}T^*=e^{-2i\theta}S^{-1}S^*$, and since α may be chosen to produce any $\theta\in[0,2\pi)$, we need only pick θ so that $-e^{2i\theta}\notin\sigma(S^{-1}S^*)$. Thus, (b) implies (c).

Next suppose that (c) holds and write $R^{-1}AR=A^*$ with $R\in M_n$ nonsingular and Hermitian. Then $R^{-1}A=A^*R^{-1}$ and $A=R(A^*R^{-1})$. But $(A^*R^{-1})^*=R^{-1}A=A^*R^{-1}$, so that A is the product of the two Hermitian matrices R and A^*R^{-1}, of which R is nonsingular, and (d) holds.

If (d) holds and $A=HK$ with H nonsingular, then $H^{-1}AH=KH=(HK)^*=A^*$, and (b) holds. The argument is similar if K is nonsingular.

Obviously (d) implies (e); we shall show that (e) implies (a). If $A=HK$ with H and K Hermitian and both singular, consider $U^*AU=(U^*HU)(U^*KU)$, where $U\in M_n$ is unitary and diagonalizes H in the form

$$U^*HU=\begin{bmatrix} D & 0\\ 0 & 0\end{bmatrix}=H'$$

for a nonsingular diagonal matrix $D\in M_k$, $k<n$. Partition the matrix U^*KU conformally with H' so that

$$U^*AU=H'(U^*KU)=\begin{bmatrix} D & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix} K' & *\\ * & *\end{bmatrix}$$

$$=\begin{bmatrix} DK' & *\\ 0 & 0\end{bmatrix}$$

The term $DK'\in M_k$ is the product of two Hermitian matrices, one of which is nonsingular, so by the equivalence of (d) and (a) it is similar to a real matrix $B\in M_k$. Denote the Jordan canonical form of B by $J\in M_k$ so that A is similar to a matrix C of the form

$$C=\begin{bmatrix} J & *\\ 0 & 0\end{bmatrix}$$

The matrix C is upper triangular, and its eigenvalues are the eigenvalues of J together with $n-k$ additional zero eigenvalues. The Jordan block structure of the Jordan canonical form of C must be the same as that of J for the blocks associated with any *nonzero* eigenvalues because if $\lambda\neq 0$, the (column) rank of each power $(C-\lambda I)^r$ is evidently equal to

$n-k+\text{rank}(J-\lambda I)^r$, $r=1,2,\ldots,n$. In particular, the Jordan blocks of C associated with any nonreal eigenvalues must occur in matching conjugate pairs, and hence the Jordan canonical form of C is similar to a matrix of the form (3.4.6), which is real. □

Problems

1. Show that every principal submatrix of a Hermitian matrix is Hermitian. Does this property hold for skew-Hermitian matrices? for normal matrices?

2. If $A\in M_n$ is Hermitian and $S\in M_n$, show that SAS^* is Hermitian. What about SAS^{-1} (if S is nonsingular)?

3. Let $A,B\in M_n$ be Hermitian. Show that A and B are similar if and only if they are unitarily similar. *Hint:* If $A=SBS^{-1}$, show that $A=U\Lambda U^*$ and $B=V\Lambda V^*$ with U and V unitary, so $U^*AU=\Lambda=V^*BV$.

4. Verify the properties 1-9 following (4.1.1).

5. Sometimes one can show that a matrix has only real eigenvalues by showing that it is similar to a Hermitian matrix. A classic example of this is the following: Let $A=[a_{ij}]\in M_n(\mathbf{R})$ be tridiagonal; that is, $a_{ij}=0$ if $|i-j|>1$. Suppose that the entries have the very weak symmetry property that $a_{i,i+1}a_{i+1,i}>0$ for all $i=1,2,\ldots,n-1$. Show that there is a real diagonal matrix D with positive diagonal entries such that DAD^{-1} is symmetric, and conclude that A has only real eigenvalues. Consider $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ and explain why the assumption on the signs of the off-diagonal entries is necessary. Use a limit argument to show that the conclusion that the eigenvalues are real continues to hold if $a_{i,i+1}a_{i+1,i}\geq 0$.

6. Show that every matrix $A\in M_n$ is uniquely determined by the Hermitian form x^*Ax that it generates, in the following sense: If $A=[a_{ij}]$, $B=[b_{ij}]\in M_n$ are given, show that $x^*Ax=x^*Bx$ for all $x\in\mathbf{C}^n$ if and only if $A=B$. *Hint:* If $x^*Ax=0$ for all $x\in\mathbf{C}^n$, consider $(x+y)^*A(x+y)$ and show that $x^*Ay+y^*Ax=0$ for all $x,y\in\mathbf{C}^n$. Choose $x=e_k$, $y=e^{i\theta}e_j$, $\theta\in\mathbf{R}$ to show that $a_{kj}e^{2i\theta}=-a_{jk}$ for all $\theta\in\mathbf{R}$ and all $j,k=1,2,\ldots,n$.

7. Show that a matrix $A\in M_n$ is not uniquely determined by the quadratic form x^TAx that it generates if $n\geq 2$; that is, if $n\geq 2$ there are $A,B\in M_n$ with $A\neq B$ such that $x^TAx=x^TBx$ for all $x\in\mathbf{C}^n$. *Hint:* If $C=-C^T$, what is x^TCx?

8. Show that a matrix $A\in M_n$ is not determined by the absolute value of the Hermitian form $|x^*Ax|$ that it generates. *Hint:* Let $A=\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ and show that $|x^*Ax|=|x^*A^Tx|$ for all $x\in\mathbf{C}^2$.

9. Show that a matrix $A\in M_n$ is almost determined by the absolute value of the Hermitian bilinear form that it generates, in the following sense: If $A,B\in M_n$ are given, show that $|x^*Ay|=|x^*By|$ for all $x,y\in\mathbf{C}^n$ if and only if $A=e^{i\theta}B$ for some $\theta\in\mathbf{R}$. *Hint:* Let $A=[a_{ij}]$ and $B=[b_{ij}]$. Use $x=e_i$ and $y=e_j$ to show $|a_{ij}|=|b_{ij}|$ for all $i,j=1,\ldots,n$. Let $x=e_i$, $y=se_j+te_k$ to show $|sa_{ij}+ta_{ik}|^2=|sb_{ij}+tb_{ik}|^2$ and hence $\text{Re}(s\bar{t}[a_{ij}\bar{a}_{ik}-b_{ij}\bar{b}_{ik}])=0$ for all $s,t\in\mathbf{C}$. Deduce that $a_{ij}/b_{ij}=a_{ik}/b_{ik}$ if $b_{ij}b_{ik}\neq 0$.

10. Show that $A\in M_n$ is Hermitian if and only if iA is skew-Hermitian. Deduce that the eigenvalues of a skew-Hermitian matrix are all pure imaginary and that the eigenvalues of the square of a skew-Hermitian matrix are real and nonpositive.

11. If $A,B\in M_n$ are Hermitian, show that $\text{tr}(AB)^2\leq\text{tr}\,A^2B^2$. *Hint:* Show that $AB-BA$ is skew-Hermitian and consider $\text{tr}(AB-BA)^2$.

12. If $A\in M_n$ is Hermitian, show that the rank of A is equal to the number of nonzero eigenvalues of A, but that this is not generally true for non-Hermitian matrices. *Hint:* $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$.

13. If $A\in M_n$ is Hermitian and $A\neq 0$, show that

$$\text{rank}(A)\geq\frac{[\text{tr}\,A]^2}{\text{tr}\,A^2}$$

with equality if and only if there is a matrix $U=[u_1\cdots u_r]\in M_{n,r}$ with orthonormal columns and some $a\in\mathbf{R}$ such that $A=aUU^*$; that is, A is a real scalar multiple of a unitary projection. *Hint:* If $\lambda_1,\ldots,\lambda_r$ are the nonzero eigenvalues of A, the Cauchy–Schwarz inequality says that

$$[\text{tr}\,A]^2=\left(\sum_{i=1}^{r}\lambda_i\right)^2\leq r\sum_{i=1}^{r}\lambda_i^2=r\,\text{tr}\,A^2$$

with equality if and only if all λ_i are equal.

14. A skew-Hermitian matrix $A\in M_n$ satisfies the identity $A=-A^*$. If $\theta\in\mathbf{R}$, show that $A=e^{i\theta}A^*$ if and only if $e^{-i\theta/2}A$ is Hermitian. What is this for $\theta=\pi$? For $\theta=0$? Explain why the class of skew-Hermitian matrices may be thought of as one of infinitely many classes of "generalized Hermitian" matrices, and describe the structure of each such class.

15. Let the Hermitian matrix $A=[a_{ij}]\in M_n$ be written in partitioned form as

$$A=\begin{bmatrix} a_{11} & x^* \\ x & \tilde{A} \end{bmatrix}$$

where $x\in\mathbf{C}^{n-1}$ and $\tilde{A}\in M_{n-1}$. Show that

$$\det A=a_{11}\det\tilde{A}-x^*(\text{adj}\,\tilde{A})x$$

where adj $\tilde{A}$ is the classical adjoint of $\tilde{A}$ [see (0.8.2)]. What weaker hypothesis on A is sufficient for this formula to hold? *Hint:* Use the Laplace cofactor expansion (0.3.1) for the determinant down the first colum of A and then across the first row of the cofactors obtained.

4.2 Variational characterizations of eigenvalues of Hermitian matrices

For a general matrix $A \in M_n$, about the only characterization of the eigenvalues is the fact that they are the roots of the characteristic equation $p_A(t)=0$. For Hermitian matrices, however, the eigenvalues can be characterized as the solutions of a series of optimization problems.

Since the eigenvalues of a Hermitian matrix $A \in M_n$ are real, we shall adopt the convention that they are labeled according to increasing (nondecreasing) size:

$$\lambda_{\min}=\lambda_1 \le \lambda_2 \le \cdots \le \lambda_{n-1} \le \lambda_n = \lambda_{\max} \tag{4.2.1}$$

The smallest and largest eigenvalues are easily characterized as the solutions to a constrained minimum and maximum problem. The characterization theorem bears the names of two British physicists, and the expression x^*Ax/x^*x is known as a *Rayleigh–Ritz ratio.*

4.2.2 **Theorem** (Rayleigh–Ritz). Let $A \in M_n$ be Hermitian, and let the eigenvalues of A be ordered as in (4.2.1). Then

$$\lambda_1 x^*x \le x^*Ax \le \lambda_n x^*x \quad \text{for all} \quad x \in \mathbf{C}^n$$

$$\lambda_{\max}=\lambda_n=\max_{x\neq 0}\frac{x^*Ax}{x^*x}=\max_{x^*x=1} x^*Ax$$

$$\lambda_{\min}=\lambda_1=\min_{x\neq 0}\frac{x^*Ax}{x^*x}=\min_{x^*x=1} x^*Ax$$

Proof: Since A is Hermitian, there is a unitary matrix $U \in M_n$ such that $A=U\Lambda U^*$ with $\Lambda=\operatorname{diag}(\lambda_1,\lambda_2,\ldots,\lambda_n)$. For any $x \in \mathbf{C}^n$ we have

$$x^*Ax=x^*U\Lambda U^*x=(U^*x)^*\Lambda(U^*x)$$

$$=\sum_{i=1}^{n}\lambda_i|(U^*x)_i|^2$$

Since each term $|(U^*x)_i|^2$ is nonnegative, we have

$$\lambda_{\min}\sum_{i=1}^{n}|(U^*x)_i|^2 \le x^*Ax=\sum_{i=1}^{n}\lambda_i|(U^*x)_i|^2 \le \lambda_{\max}\sum_{i=1}^{n}|(U^*x)_i|^2$$

Because U is unitary,

$$\sum_{i=1}^{n}|(U^*x)_i|^2=\sum_{i=1}^{n}|x_i|^2=x^*x$$

and hence we have shown that

$$\lambda_1 x^*x=\lambda_{\min}\, x^*x \le x^*Ax \le \lambda_{\max}\, x^*x=\lambda_n x^*x \tag{4.2.3}$$

These inequalities are sharp, for if x is an eigenvector of A corresponding to the eigenvalue λ_1, then $x^*Ax=x^*\lambda_1 x=\lambda_1 x^*x$, and similarly for λ_n. The remaining assertions follow easily from (4.2.3). If $x\neq 0$, we have

$$\frac{x^*Ax}{x^*x}\le \lambda_n$$

with equality when x is a λ_n eigenvector of A, so

$$\max_{x\neq 0}\frac{x^*Ax}{x^*x}=\lambda_n \tag{4.2.4}$$

Finally, if $x\neq 0$, we have

$$\frac{x^*Ax}{x^*x}=\left(\frac{x}{\sqrt{x^*x}}\right)^*A\left(\frac{x}{\sqrt{x^*x}}\right) \quad \text{and} \quad \left(\frac{x}{\sqrt{x^*x}}\right)^*\left(\frac{x}{\sqrt{x^*x}}\right)=1$$

so the condition (4.2.4) is equivalent to the condition

$$\max_{x^*x=1} x^*Ax=\lambda_n \tag{4.2.5}$$

The arguments for λ_1 are similar. □

The geometrical interpretation of (4.2.5) is that λ_n is the largest value of the function x^*Ax as x ranges over the unit sphere in $\mathbf{C}^n$, a compact set. The bounds (4.2.3) give the following *eigenvalue inclusion* result.

4.2.6 **Corollary.** Let $A \in M_n$ be a given Hermitian matrix, let $x \in \mathbf{C}^n$ be a given nonzero vector, and let $\alpha \equiv x^*Ax/x^*x$. Then there is at least one eigenvalue of A in the interval $(-\infty,\alpha]$ and at least one in $[\alpha,\infty)$.

Exercise. Prove Corollary (4.2.6).

Exercise. What are the analog and the geometrical interpretation of (4.2.5) for λ_1?

The Rayleigh–Ritz theorem provides a variational characterization of the largest and smallest eigenvalues of a Hermitian matrix A, but what about the rest of the eigenvalues? Suppose $A=U\Lambda U^*$ with $U=[u_1\ u_2\ \ldots\ u_n]$; the columns of U are the orthonormal eigenvectors of A. If we consider only those vectors $x \in \mathbf{C}^n$ that are orthogonal to u_1, we have the following modification in the main identity of (4.2.2):

$$x^*Ax=\sum_{i=1}^n \lambda_i|(U^*x)_i|^2=\sum_{i=1}^n \lambda_i|u_i^*x|^2=\sum_{i=2}^n \lambda_i|u_i^*x|^2$$

This is a nonnegative linear combination of $\lambda_2, \lambda_3, \ldots, \lambda_n$, and hence

$$x^*Ax=\sum_{i=2}^n \lambda_i|u_i^*x|^2 \geq \lambda_2 \sum_{i=2}^n |u_i^*x|^2=\lambda_2 \sum_{i=1}^n |(U^*x)_i|^2=\lambda_2 x^*x$$

provided that x is orthogonal to the first column of U. This inequality becomes an equality if we choose $x=u_2$, so we have a characterization

$$\min_{\substack{x\neq 0\\ x\perp u_1}} \frac{x^*Ax}{x^*x}=\min_{\substack{x^*x=1\\ x\perp u_1}} x^*Ax=\lambda_2 \tag{4.2.7}$$

of the second smallest eigenvalue.

Exercise. Extend this argument to show that

$$\min_{\substack{x\neq 0\\ x\perp u_1, u_2,\ldots,u_{k-1}}} \frac{x^*Ax}{x^*x}=\min_{\substack{x^*x=1\\ x\perp u_1, u_2,\ldots,u_{k-1}}} x^*Ax=\lambda_k, \quad k=2,3,\ldots,n \tag{4.2.8}$$

Exercise. Show that

$$\max_{\substack{x\neq 0\\ x\perp u_n, u_{n-1},\ldots,u_{n-k+1}}} \frac{x^*Ax}{x^*x}=\max_{\substack{x^*x=1\\ x\perp u_n, u_{n-1},\ldots,u_{n-k+1}}} x^*Ax=\lambda_{n-k}, \quad k=1,2,\ldots,n-1 \tag{4.2.9}$$

Unfortunately, these formulae are of little practical value because they require explicit knowledge of some of the eigenvectors, about which we usually have no information. But (4.2.7) and the related formulae (4.2.8) and (4.2.9) can be the starting point for developing a useful characterization. Let $w\in \mathbf{C}^n$ be a given vector. Then

$$\begin{aligned}
\sup_{\substack{x^*x=1\\ x\perp w}} x^*Ax &= \sup_{\substack{x^*x=1\\ x\perp w}} x^*U\Lambda U^*x=\sup_{\substack{x^*x=1\\ x\perp w}} \sum_{i=1}^n \lambda_i|(U^*x)_i|^2\\
&= \sup_{\substack{z^*z=1\\ x=Uz\perp w}} \sum_{i=1}^n \lambda_i|z_i|^2=\sup_{\substack{z^*z=1\\ z\perp U^*w}} \sum_{i=1}^n \lambda_i|z_i|^2\\
&\geq \sup_{\substack{z^*z=1\\ z\perp U^*w\\ z_1=z_2=\cdots=z_{n-2}=0}} \sum_{i=1}^n \lambda_i|z_i|^2\\
&= \sup_{\substack{|z_{n-1}|^2+|z_n|^2=1\\ z\perp U^*w}} \lambda_{n-1}|z_{n-1}|^2+\lambda_n|z_n|^2 \geq \lambda_{n-1}
\end{aligned} \tag{4.2.10}$$

In the second line of this argument we set $z\equiv U^*x$ and used the fact that U is unitary to conclude that $z^*z=1$ if $x^*x=1$. The first inequality in the last line comes from the fact that if one restricts the set over which a supremum is taken, the value of the supremum cannot increase. The final inequality follows from the ordering $\lambda_n \geq \lambda_{n-1}$ and an often used property of convex combinations.

In the foregoing argument, the vector w was fixed but arbitrary, and hence we may take the infimum of (4.2.10) over all w to obtain

$$\inf_{w\in \mathbf{C}^n} \sup_{\substack{x^*x=1\\ x\perp w}} x^*Ax \geq \lambda_{n-1}$$

We have seen, however, that equality holds in (4.2.10) if $w=u_n$, so we have

$$\inf_{w\in \mathbf{C}^n} \sup_{\substack{x^*x=1\\ x\perp w}} x^*Ax = \lambda_{n-1}$$

a characterization that is somewhat more complicated in form than (4.2.7) but does not involve knowledge of any of the eigenvectors of A. Since $x^*Ax=\lambda_{n-1}$ for $x=u_{n-1}$, one often sees this formula with "max" instead of "sup" and with "min" instead of "inf." This is the basic idea behind the following Courant–Fischer "min-max theorem".

4.2.11 Theorem (Courant–Fischer). Let $A\in M_n$ be a Hermitian matrix with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, and let k be a given integer with $1\leq k\leq n$. Then

$$\min_{w_1, w_2,\ldots,w_{n-k}\in \mathbf{C}^n} \max_{\substack{x\neq 0,\, x\in \mathbf{C}^n\\ x\perp w_1, w_2,\ldots,w_{n-k}}} \frac{x^*Ax}{x^*x}=\lambda_k \tag{4.2.12}$$

and

$$\max_{w_1, w_2,\ldots,w_{k-1}\in \mathbf{C}^n} \min_{\substack{x\neq 0,\, x\in \mathbf{C}^n\\ x\perp w_1, w_2,\ldots,w_{k-1}}} \frac{x^*Ax}{x^*x}=\lambda_k \tag{4.2.13}$$

Remark: If $k=n$ in (4.2.12) or $k=1$ in (4.2.13), we agree to omit the outer optimization, as the set over which the optimization takes place is empty. In these two cases the assertions reduce to the Rayleigh–Ritz theorem (4.2.2).

Proof: We consider only (4.2.12), as the argument for (4.2.13) is similar. Write $A=U\Lambda U^*$ with U unitary and $\Lambda=\text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, and let $1<k\leq n$. If $x\neq 0$,

$$\frac{x^*Ax}{x^*x}=\frac{(U^*x)^*\Lambda(U^*x)}{x^*x}=\frac{(U^*x)^*\Lambda(U^*x)}{(U^*x)^*(U^*x)}$$

and $\{U^*x : x \in \mathbf{C}^n \text{ and } x \neq 0\} = \{y \in \mathbf{C}^n : y \neq 0\}$. Thus, if $w_1, w_2, \ldots, w_{n-k} \in \mathbf{C}^n$ are given, we have

$$\begin{aligned}
\sup_{\substack{x \neq 0 \\ x \perp w_1, \ldots, w_{n-k}}} \frac{x^*Ax}{x^*x} &= \sup_{\substack{y \neq 0 \\ y \perp U^*w_1, \ldots, U^*w_{n-k}}} \frac{y^*\Lambda y}{y^*y} \\
&= \sup_{\substack{y^*y = 1 \\ y \perp U^*w_1, \ldots, U^*w_{n-k}}} \sum_{i=1}^{n} \lambda_i |y_i|^2 \\
&\geq \sup_{\substack{y^*y = 1 \\ y \perp U^*w_1, \ldots, U^*w_{n-k} \\ y_1 = y_2 = \cdots = y_{k-1} = 0}} \sum_{i=1}^{n} \lambda_i |y_i|^2 \\
&= \sup_{\substack{|y_k|^2 + |y_{k+1}|^2 + \cdots + |y_n|^2 = 1 \\ y \perp U^*w_1, \ldots, U^*w_{n-k}}} \sum_{i=k}^{n} \lambda_i |y_i|^2 \geq \lambda_k
\end{aligned}$$

This shows that

$$\sup_{\substack{x \neq 0 \\ x \perp w_1, \ldots, w_{n-k}}} \frac{x^*Ax}{x^*x} \geq \lambda_k$$

for any $n-k$ vectors $w_1, \ldots, w_{n-k}$. But (4.2.9) shows that equality holds for one choice of the vectors w_i, namely $w_i = u_{n-i+1}$, where $U = [u_1 \ldots u_n]$. Therefore,

$$\inf_{w_1, \ldots, w_{n-k}} \sup_{\substack{x \neq 0 \\ x \perp w_1, \ldots, w_{n-k}}} \frac{x^*Ax}{x^*x} = \lambda_k$$

and we may replace "inf" and "sup" with "min" and "max" since the extremum is achieved. The argument for (4.2.13) is similar. □

Exercise. Provide the details for a proof of (4.2.13).

Problems

1. Let $A \in M_n$ be a Hermitian matrix with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Use (4.2.11) to show that

$$\lambda_k = \min_{S_k \subset \mathbf{C}^n} \max_{\substack{x \neq 0 \\ x \in S_k}} \frac{x^*Ax}{x^*x}, \qquad k = 1, 2, \ldots, n$$

$$\lambda_k = \max_{S_{n-k+1} \subset \mathbf{C}^n} \min_{\substack{x \neq 0 \\ x \in S_{n-k+1}}} \frac{x^*Ax}{x^*x}, \qquad k = 1, 2, \ldots, n$$

where, in both cases, S_j denotes a subspace of dimension j and the outer optimization is over all subspaces of the indicated dimension.

2. If $A \in M_n$ is Hermitian, show that the following three optimization problems all have the same solution:

(a) $\displaystyle\max_{x^*x = 1} x^*Ax$

(b) $\displaystyle\max_{x \neq 0} \frac{x^*Ax}{x^*x}$

(c) $\displaystyle\max_{x^*Ax = 1} \frac{1}{x^*x}$ if at least one eigenvalue of A is positive

3. If $A \in M_n$ is Hermitian and $x^*x = 1$, show that

$$\lambda_{\max} \geq x^*Ax \geq \lambda_{\min}$$

4. Show that the assumption that A is Hermitian is essential in (4.2.2) by considering $A = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$. What is $\max\{x^TAx/x^Tx : 0 \neq x \in \mathbf{R}^2\}$? What is $\max \operatorname{Re}\{x^*Ax/x^*x : 0 \neq x \in \mathbf{C}^2\}$?

5. Let $A \in M_n$ have eigenvalues $\{\lambda_i\}$. Show that, even if A is not Hermitian, one has the bounds

$$\min_{x \neq 0} \left| \frac{x^*Ax}{x^*x} \right| \leq |\lambda_i| \leq \max_{x \neq 0} \left| \frac{x^*Ax}{x^*x} \right|, \qquad i = 1, 2, \ldots, n$$

Hint: Consider $x =$ an eigenvector of A, and $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ to show that neither bound need be sharp.

4.3 Some applications of the variational characterizations

Among the many important applications of the Courant–Fischer theorem, one of the simplest is to the problem of comparing the eigenvalues of $A+B$ with those of A. We denote the eigenvalues of a matrix A by $\lambda_i(A)$.

4.3.1 **Theorem** (Weyl). Let $A, B \in M_n$ be Hermitian and let the eigenvalues $\lambda_i(A)$, $\lambda_i(B)$, and $\lambda_i(A+B)$ be arranged in increasing order (4.2.1). For each $k = 1, 2, \ldots, n$ we have

$$\lambda_k(A) + \lambda_1(B) \leq \lambda_k(A+B) \leq \lambda_k(A) + \lambda_n(B) \qquad (4.3.2)$$

Proof: For any nonzero $x \in \mathbf{C}^n$ we have the bound

$$\lambda_1(B) \leq \frac{x^*Bx}{x^*x} \leq \lambda_n(B)$$

and hence for any $k = 1, 2, \ldots, n$ we have

$$\lambda_k(A+B) = \min_{w_1,\ldots,w_{n-k}\in \mathbf{C}^n} \max_{\substack{x\neq 0\\ x\perp w_1,\ldots,w_{n-k}}} \frac{x^*(A+B)x}{x^*x}$$

$$= \min_{w_1,\ldots,w_{n-k}\in \mathbf{C}^n} \max_{\substack{x\neq 0\\ x\perp w_1,\ldots,w_{n-k}}} \left[\frac{x^*Ax}{x^*x}+\frac{x^*Bx}{x^*x}\right]$$

$$\geq \min_{w_1,\ldots,w_{n-k}\in \mathbf{C}^n} \max_{\substack{x\neq 0\\ x\perp w_1,\ldots,w_{n-k}}} \left[\frac{x^*Ax}{x^*x}+\lambda_1(B)\right] = \lambda_k(A)+\lambda_1(B)$$

A similar argument establishes the upper bound as well. □

Exercise. Show that equality may be attained in the bounds given in (4.3.1). *Hint:* Let $\{u_1, u_2, \ldots, u_n\}$ be an orthonormal set of eigenvectors of A with $Au_i = \lambda_i(A)u_i$. Consider $B = \alpha u_i u_i^*$ for $\alpha > 0$ and for $\alpha < 0$.

Weyl's theorem gives two-sided bounds for the eigenvalues of $A+B$ for *any* Hermitian matrices A and B. Further refinements can be obtained by restricting B to have a special form – for example, positive definite, rank 1, rank k, or a bordering matrix.

A matrix $B\in M_n$ such that $x^*Bx\geq 0$ for all $x\in \mathbf{C}^n$ is said to be *positive semidefinite*; an equivalent condition is that B be Hermitian and have all eigenvalues nonnegative (see Chapter 7). The following result, an immediate corollary of Weyl's theorem known as the *monotonicity theorem*, says that all the eigenvalues of a Hermitian matrix increase if a positive semidefinite matrix is added to it.

4.3.3 **Corollary.** Let $A, B\in M_n$ be Hermitian. Assume that B is positive semidefinite and that the eigenvalues of A and $A+B$ are arranged in increasing order (4.2.1). Then

$$\lambda_k(A)\leq \lambda_k(A+B) \quad \text{for all} \quad k=1,2,\ldots,n$$

Proof: Use the lower bound in (4.3.2) and the fact that $\lambda_1(B)\geq 0$. □

If the matrix B is of rank 1, the bounds on the eigenvalues of $A+B$ compared with those of A are in the form of an *interlacing theorem:* between each successive pair of odd-numbered (or even-numbered) eigenvalues of $A+B$ there is at least one eigenvalue of A.

4.3.4 **Theorem.** Let $A\in M_n$ be Hermitian and let $z\in \mathbf{C}^n$ be a given vector. If the eigenvalues of A and $A\pm zz^*$ are arranged in increasing order (4.2.1), we have

(a) $\lambda_k(A\pm zz^*)\leq \lambda_{k+1}(A)\leq \lambda_{k+2}(A\pm zz^*), \quad k=1,2,\ldots,n-2$

(b) $\lambda_k(A)\leq \lambda_{k+1}(A\pm zz^*)\leq \lambda_{k+2}(A), \quad k=1,2,\ldots,n-2$

Proof: Let $1\leq k\leq n-2$ and use (4.2.12) to write

$$\lambda_{k+2}(A\pm zz^*) = \min_{w_1,\ldots,w_{n-k-2}\in \mathbf{C}^n} \max_{\substack{x\neq 0\\ x\perp w_1,\ldots,w_{n-k-2}}} \frac{x^*(A\pm zz^*)x}{x^*x}$$

$$\geq \min_{w_1,\ldots,w_{n-k-2}\in \mathbf{C}^n} \max_{\substack{x\neq 0\\ x\perp w_1,\ldots,w_{n-k-2}\\ x\perp z}} \frac{x^*(A\pm zz^*)x}{x^*x}$$

$$= \min_{\substack{w_1,\ldots,w_{n-k-2}\in \mathbf{C}^n\\ w_{n-k-1}\equiv z}} \max_{\substack{x\neq 0\\ x\perp w_1,\ldots,w_{n-k-2},w_{n-k-1}}} \frac{x^*Ax}{x^*x}$$

$$\geq \min_{w_1,\ldots,w_{n-k-2},w_{n-k-1}\in \mathbf{C}^n} \max_{\substack{x\neq 0\\ x\perp w_1,\ldots,w_{n-k-2},w_{n-k-1}}} \frac{x^*Ax}{x^*x}$$

$$= \lambda_{k+1}(A)$$

Now let $2\leq k\leq n-1$ and use (4.2.13) to write

$$\lambda_k(A\pm zz^*) = \max_{w_1,\ldots,w_{k-1}\in \mathbf{C}^n} \min_{\substack{x\neq 0\\ x\perp w_1,\ldots,w_{k-1}}} \frac{x^*(A\pm zz^*)x}{x^*x}$$

$$\leq \max_{w_1,\ldots,w_{k-1}\in \mathbf{C}} \min_{\substack{x\neq 0\\ x\perp w_1,\ldots,w_{k-1}\\ x\perp z}} \frac{x^*(A\pm zz^*)x}{x^*x}$$

$$= \max_{\substack{w_1,\ldots,w_{k-1}\in \mathbf{C}^n\\ w_k\equiv z}} \min_{\substack{x\neq 0\\ x\perp w_1,\ldots,w_{k-1},w_k}} \frac{x^*Ax}{x^*x}$$

$$\leq \max_{w_1,\ldots,w_{k-1},w_k\in \mathbf{C}^n} \min_{\substack{x\neq 0\\ x\perp w_1,\ldots,w_{k-1},w_k}} \frac{x^*Ax}{x^*x} = \lambda_{k+1}(A)$$

Taken together, these two families of inequalities yield the asserted inequalities. □

If $B\in M_n$ is a Hermitian matrix, and if $B = U\Lambda U^*$ with $U = [u_1\ u_2 \cdots u_n]$ unitary and $\Lambda = \operatorname{diag}(\beta_1, \beta_2, \ldots, \beta_n)$, then the rank of B is equal to the number of nonzero eigenvalues. If B has rank less than or equal to r, then we may assume that $\beta_{r+1} = \cdots = \beta_n = 0$. If the rank is less than r, then some of $\beta_1, \beta_2, \ldots, \beta_r$ will be zero as well. The expression

$$B=\sum_{i=1}^{r}\beta_i u_i u_i^* \qquad (4.3.5)$$

is another way to write $B=U\Lambda U^*$. Conversely, any matrix of the form (4.3.5), with all $\beta_i\neq 0$ and $\{u_i\}$ an independent set, has rank r; if the u_i terms are not known to be independent, then the rank of B is at most r. The next result, a theorem of Weyl that has its origins in the theory of integral equations, gives bounds for the eigenvalues of $A+B$ when B has rank r. It is an easy generalization of the rank 1 case (4.3.4).

4.3.6 **Theorem.** Let $A,B\in M_n$ be Hermitian and suppose that B has rank at most r. Then

(a) $\lambda_k(A+B)\leq\lambda_{k+r}(A)\leq\lambda_{k+2r}(A+B), \quad k=1,2,\ldots,n-2r$

(b) $\lambda_k(A)\leq\lambda_{k+r}(A+B)\leq\lambda_{k+2r}(A), \quad k=1,2,\ldots,n-2r$

(c) If $A=U\Lambda U^*$ with $U=[u_1\; u_2\;\cdots\; u_n]\in M_n$ unitary and $\Lambda=\operatorname{diag}(\lambda_1,\ldots,\lambda_n)$ with $\lambda_1\leq\lambda_2\leq\cdots\leq\lambda_n$, and if

$$B=\lambda_n u_n u_n^*+\lambda_{n-1}u_{n-1}u_{n-1}^*+\cdots+\lambda_{n-r+1}u_{n-r+1}u_{n-r+1}^*$$

then $\lambda_{\max}(A-B)=\lambda_{n-r}(A)$.

Proof: Let $B=\alpha_1 u_1 u_1^*+\cdots+\alpha_r u_r u_r^*$, where the set $\{u_1,\ldots,u_r\}\subset\mathbf{C}^n$ is not necessarily independent. The proofs of (a) and (b) are exactly the same as those of (a) and (b) in (4.3.4), except that where one imposed the single condition "$x\perp z$" before, one now imposes the r conditions "$x\perp u_1,\ldots,u_r$" and completes the argument accordingly. For (c), observe that $u_1,\ldots,u_n$ are all eigenvectors of $A-B$, but $(A-B)u_k=0$ for $k=n-r+1,n-r+2,\ldots,n$ and $(A-B)u_k=\lambda_k u_k$ for $k=1,2,\ldots,n-r$. Since $\lambda_{n-r}\geq\lambda_{n-r-1}\geq\cdots\geq\lambda_1$, the largest eigenvalue of $A-B$ is λ_{n-r}. □

Exercise. Provide the details for the proofs of (a) and (b) in (4.3.6).

The preceding result gives us enough information to derive the following general result of Weyl on the eigenvalues of a sum of Hermitian matrices.

4.3.7 **Theorem** (Weyl). Let $A,B\in M_n$ be Hermitian matrices, and let the eigenvalues of A,B and $A+B$ be arranged in increasing order (4.2.1). Then for every pair of integers j,k such that $1\leq j,k\leq n$ and $j+k\geq n+1$ we have

$$\lambda_{j+k-n}(A+B)\leq\lambda_j(A)+\lambda_k(B)$$

and for every pair of integers j,k such that $1\leq j,k\leq n$ and $j+k\leq n+1$ we have

$$\lambda_j(A)+\lambda_k(B)\leq\lambda_{j+k-1}(A+B)$$

Proof: Let j,k be given integers satisfying the first set of conditions. Let $A=U\Lambda(A)U^*$ and $B=V\Lambda(B)V^*$, where $U=[u_1\; u_2\;\cdots\; u_n]\in M_n$ and $V=[v_1\; v_2\;\cdots\; v_n]\in M_n$ are unitary, $\Lambda(A)=\operatorname{diag}(\lambda_1(A),\ldots,\lambda_n(A))\in M_n$, and $\Lambda(B)=\operatorname{diag}(\lambda_1(B),\ldots,\lambda_n(B))\in M_n$. Then

$$A_j\equiv\lambda_n(A)u_n u_n^*+\lambda_{n-1}(A)u_{n-1}u_{n-1}^*+\cdots+\lambda_{j+1}(A)u_{j+1}u_{j+1}^*$$

has rank at most $n-j$, $B_k\equiv\lambda_n(B)v_n v_n^*+\cdots+\lambda_{k+1}(B)v_{k+1}v_{k+1}^*$ has rank at most $n-k$, and A_j+B_k has rank at most $2n-j-k$. Then $\lambda_n(A-A_j)=\lambda_j(A)$ and $\lambda_n(B-B_k)=\lambda_k(B)$ by (4.3.6c), and $\lambda_n(A-A_j+B-B_k)=\lambda_n(A+B-(A_j+B_k))\geq\lambda_{n-(2n-j-k)}(A+B)=\lambda_{j+k-n}(A+B)$ by (4.3.6b) (with $k+r=n$ and $r=2n-j-k$). Also,

$$\lambda_n(A-A_j+B-B_k)\leq\lambda_n(A-A_j)+\lambda_n(B-B_k)$$

by (4.3.2) (with $k=n$). Thus, we have

$$\lambda_j(A)+\lambda_k(B)=\lambda_n(A-A_j)+\lambda_n(B-B_k)\geq\lambda_n(A-A_j+B-B_k)$$
$$=\lambda_n((A+B)-(A_j+B_k))\geq\lambda_{j+k-n}(A+B)$$

which is the first asserted inequality. The second set of inequalities follows directly from the first set when it is applied to $-A$ and $-B$. □

Exercise. Provide the details for the deduction of the second set of inequalities in (4.3.7) from the first set. *Hint:* Apply (4.3.7) to obtain an upper bound for $\lambda_{j+k-n}(-A-B)$ and use the fact that $\lambda_i(-A)=-\lambda_{n-i+1}(A)$ if $A\in M_n$ is Hermitian.

As a final result of this type, we consider an interlacing theorem for the eigenvalues of $A+B$, where each of A and B is assumed to have a special form. The result, known as the *interlacing eigenvalues theorem for bordered matrices,* is similar to the case (4.3.4) in which B is assumed to have rank 1.

4.3.8 **Theorem.** Let $A\in M_n$ be a given Hermitian matrix, let $y\in\mathbf{C}^n$ be a given vector, and let $a\in\mathbf{R}$ be a given real number. Let $\hat{A}\in M_{n+1}$ be the Hermitian matrix obtained by bordering A with y and a as follows:

$$\hat{A}\equiv\left[\begin{array}{c|c} A & y \\ \hline y^* & a \end{array}\right]$$

Let the eigenvalues of A and $\hat{A}$ be denoted by $\{\lambda_i\}$ and $\{\hat{\lambda}_i\}$, respectively, and assume that they have been arranged in increasing order $\lambda_1 \le \cdots \le \lambda_n$ and $\hat{\lambda}_1 \le \hat{\lambda}_2 \le \cdots \le \hat{\lambda}_n \le \hat{\lambda}_{n+1}$. Then

$$\hat{\lambda}_1 \le \lambda_1 \le \hat{\lambda}_2 \le \lambda_2 \le \cdots \le \lambda_{n-1} \le \hat{\lambda}_n \le \lambda_n \le \hat{\lambda}_{n+1} \tag{4.3.9}$$

Proof: Let an integer k be given with $1 \le k \le n$. We shall prove that $\hat{\lambda}_k \le \lambda_k \le \hat{\lambda}_{k+1}$. Let $\hat{x} = [x^T \xi]^T \in \mathbf{C}^{n+1}$, $x \in \mathbf{C}^n$, $\xi \in \mathbf{C}$, and $\hat{w}_i = [w_i^T \omega]^T \in \mathbf{C}^{n+1}$, $w_i \in \mathbf{C}^n$, $\omega \in \mathbf{C}$. Use (4.2.12) of the Courant–Fischer theorem to write

$$\begin{aligned}
\hat{\lambda}_{k+1} &= \min_{\hat{w}_1,\ldots,\hat{w}_{(n+1)-(k+1)} \in \mathbf{C}^{n+1}} \ \max_{\substack{\hat{x} \ne 0,\ \hat{x} \in \mathbf{C}^{n+1} \\ \hat{x} \perp \hat{w}_1,\ldots,\hat{w}_{(n+1)-(k+1)}}} \frac{\hat{x}^* \hat{A} \hat{x}}{\hat{x}^* \hat{x}} \\
&\ge \min_{\hat{w}_1,\ldots,\hat{w}_{n-k} \in \mathbf{C}^{n+1}} \ \max_{\substack{\hat{x} \ne 0 \\ \hat{x} \perp \hat{w}_1,\ldots,\hat{w}_{n-k} \\ \hat{x} \perp e_{n+1}}} \frac{\hat{x}^* \hat{A} \hat{x}}{\hat{x}^* \hat{x}} \\
&= \min_{w_1,\ldots,w_{n-k} \in \mathbf{C}^n} \ \max_{\substack{x \ne 0,\ x \in \mathbf{C}^n \\ x \perp w_1,\ldots,w_{n-k}}} \frac{x^* A x}{x^* x} = \lambda_k
\end{aligned}$$

For the lower bound on λ_k we use (4.2.13).

$$\begin{aligned}
\hat{\lambda}_k &= \max_{\hat{w}_1,\ldots,\hat{w}_{k-1} \in \mathbf{C}^{n+1}} \ \min_{\substack{\hat{x} \ne 0,\ \hat{x} \in \mathbf{C}^{n+1} \\ \hat{x} \perp \hat{w}_1,\ldots,\hat{w}_{k-1}}} \frac{\hat{x}^* \hat{A} \hat{x}}{\hat{x}^* \hat{x}} \\
&\le \max_{\hat{w}_1,\ldots,\hat{w}_{k-1} \in \mathbf{C}^{n+1}} \ \min_{\substack{\hat{x} \ne 0 \\ \hat{x} \perp \hat{w}_1,\ldots,\hat{w}_{k-1} \\ \hat{x} \perp e_{n+1}}} \frac{\hat{x}^* \hat{A} \hat{x}}{\hat{x}^* \hat{x}} \\
&= \max_{w_1,\ldots,w_{k-1} \in \mathbf{C}^n} \ \min_{\substack{x \ne 0,\ x \in \mathbf{C}^n \\ x \perp w_1,\ldots,w_{k-1}}} \frac{x^* A x}{x^* x} = \lambda_k
\end{aligned}$$

□

We have seen two examples of interlacing theorems for eigenvalues: if a given Hermitian matrix is modified by adding a rank one Hermitian matrix or by bordering, then the new and old eigenvalues must interlace. What about the converse? If two interlacing sets of real numbers are given, can they be realized by a Hermitian matrix and a suitable modification? The answer is in the affirmative, and we give as an example a converse to Theorem (4.3.8).

4.3.10 Theorem. Let n be a given positive integer, and let

$$\{\lambda_i : i = 1, 2, \ldots, n\} \quad \text{and} \quad \{\hat{\lambda}_i : i = 1, 2, \ldots, n, n+1\}$$

be two given sequences of real numbers such that

$$\hat{\lambda}_1 \le \lambda_1 \le \hat{\lambda}_2 \le \lambda_2 \le \cdots \le \lambda_{n-1} \le \hat{\lambda}_n \le \lambda_n \le \hat{\lambda}_{n+1}$$

Let $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. There exists a real number a and a real vector $y \in \mathbf{R}^n$ such that $\{\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_{n+1}\}$ is the set of eigenvalues of the real symmetric matrix

$$\hat{A} \equiv \begin{bmatrix} \Lambda & y \\ y^T & a \end{bmatrix} \in M_{n+1}(\mathbf{R})$$

Proof: Obviously $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ is the set of eigenvalues of Λ, and since $\operatorname{tr} \hat{A} = \operatorname{tr} \Lambda + a$, we must have $a = \operatorname{tr} \hat{A} - \operatorname{tr} \Lambda = \sum_{i=1}^{n+1} \hat{\lambda}_i - \sum_{i=1}^{n} \lambda_i$. The characteristic polynomial $p_{\hat{A}}(t)$ of $\hat{A}$ is easily computed as

$$\begin{aligned}
&\det(tI - \hat{A}) \\
&= \det \begin{bmatrix} tI - \Lambda & -y \\ -y^T & t-a \end{bmatrix} \\
&= \det \begin{bmatrix} I & 0 \\ [(tI-\Lambda)^{-1} y]^T & 1 \end{bmatrix} \begin{bmatrix} tI - \Lambda & -y \\ -y^T & t-a \end{bmatrix} \begin{bmatrix} I & (tI-\Lambda)^{-1} y \\ 0 & 1 \end{bmatrix} \\
&= \det \begin{bmatrix} tI - \Lambda & 0 \\ 0 & (t-a) - y^T (tI-\Lambda)^{-1} y \end{bmatrix} \\
&= [(t-a) - y^T (tI-\Lambda)^{-1} y] \det(tI - \Lambda) \\
&= \left[(t-a) - \sum_{i=1}^{n} y_i^2 \frac{1}{t - \lambda_i} \right] \prod_{i=1}^{n} (t - \lambda_i) = p_{\hat{A}}(t)
\end{aligned} \tag{4.3.11}$$

We have already determined the necessary value of a, so it remains to be shown that n real numbers y_i can be found for (4.3.11) so that $p_{\hat{A}}(\hat{\lambda}_k) = 0$ for $k = 1, 2, \ldots, n+1$.

Define the polynomials

$$f(t) = \prod_{i=1}^{n+1} (t - \hat{\lambda}_i), \qquad \text{degree of } f = n+1 \tag{4.3.12}$$

$$g(t) = \prod_{i=1}^{n} (t - \lambda_i), \qquad \text{degree of } g = n \tag{4.3.13}$$

By the Euclidean algorithm we must have

$$f(t) = g(t)(t-c) + r(t)$$

where c is a real number and $r(t)$ is a polynomial of degree at most $n-1$. By explicit computation we find that $c = \sum_{i=1}^{n+1} \hat{\lambda}_i - \sum_{i=1}^{n} \lambda_i = a$. Furthermore, $f(\lambda_k) = g(\lambda_k)(\lambda_k - a) + r(\lambda_k) = r(\lambda_k)$ for $k = 1, 2, \ldots, n$

because $g(\lambda_k)=0$. The polynomial $r(t)$ is therefore known at n points and can be written explicitly in terms of Lagrange interpolating polynomials if the points of interpolation $\lambda_1,\ldots,\lambda_n$ are distinct. Under this assumption, $g(t)$ has only simple roots, and the Lagrange interpolation formula for $r(t)$ is

$$r(t)=\sum_{i=1}^{n} f(\lambda_i)\frac{g(t)}{g'(\lambda_i)(t-\lambda_i)}$$

Thus,

$$\frac{f(t)}{g(t)}=(t-a)+\frac{r(t)}{g(t)}=(t-a)-\sum_{i=1}^{n}\frac{-f(\lambda_i)}{g'(\lambda_i)}\frac{1}{t-\lambda_i}$$

Because $f(\hat\lambda_k)=0$ for all $k=1,2,\ldots,n+1$ we must have

$$(\hat\lambda_k-a)-\sum_{i=1}^{n}\frac{-f(\lambda_i)}{g'(\lambda_i)}\frac{1}{\hat\lambda_k-\lambda_i}=0,\qquad k=1,2,\ldots,n+1 \tag{4.3.14}$$

Notice that if $\hat\lambda_k=\lambda_i$ for $i=k-1$ or k, then the corresponding term $1/(t-\lambda_i)$ has a zero coefficient and there is no singularity at $t=\hat\lambda_k$. If we can set $y_i^2\equiv -f(\lambda_i)/g'(\lambda_i)$ for $i=1,2,\ldots,n$, then (4.3.11) guarantees that $p_{\hat A}(\hat\lambda_k)=0$ and we are done. We must show, therefore, that $f(\lambda_i)/g'(\lambda_i)\le 0$ for $i=1,2,\ldots,n$, and it is now that the interlacing assumption must be used. Using the definitions of $f(t)$ and $g(t)$ and the interlacing assumption, we find that

$$f(\lambda_i)=(-1)^{n-i+1}\prod_{j=1}^{n+1}|\lambda_i-\hat\lambda_j|$$

$$g'(\lambda_i)=(-1)^{n-i}\prod_{\substack{j=1\\ j\ne i}}^{n}|\lambda_i-\lambda_j|$$

and hence $f(\lambda_i)$ and $g'(\lambda_i)$ always have opposite signs.

It is easy to modify the argument to cover the case in which some of the λ_i terms coincide. If, for example, $\lambda_1=\lambda_2=\cdots=\lambda_k<\lambda_{k+1}\le\cdots$ for some $k\ge 2$, then $\hat\lambda_2=\cdots=\hat\lambda_k=\lambda_1$. The polynomial $f(t)$ in (4.3.12) has a factor $(t-\hat\lambda_1)(t-\lambda_1)^{k-1}$; the polynomial $g(t)$ in (4.3.13) has a factor $(t-\lambda_1)^k$ and k is the exact multiplicity of λ_1 as a zero of $g(t)$. We may therefore modify $f(t)$, $g(t)$, and $r(t)$ by dividing each by $(t-\lambda_1)^{k-1}$. The modified polynomial $g(t)$ will have λ_1 as a simple zero. If we proceed in this way to remove all the multiple roots of $g(t)$, the argument can proceed as before, and the conclusion is the same. □

The preceding results treat the situation in which a Hermitian matrix is "bordered" by adding a new last row and column, but they could also be thought of as giving information about the behavior of the eigenvalues

of a Hermitian matrix when its last row and column are *deleted*. There is, of course, nothing special about the *last* row and column. If the ith row and column of the matrix $\hat A$ in (4.3.8) are deleted instead of the $(n+1)$st, one merely changes e_{n+1} to e_i in the proof and obtains the same interlacing inequalities (4.3.9).

Theorems (4.3.8) and (4.3.10) together say that the interlacing inequalities (4.3.9) are a complete description of the relationship between the eigenvalues of a Hermitian matrix and the eigenvalues of any one of its principal submatrices of order $n-1$. If one considers simultaneously all n of the $(n-1)$-by-$(n-1)$ principal submatrices of A, more can be said. Let A_j denote the principal submatrix obtained by deleting the jth row and column of A, $j=1,2,\ldots,n$, and let the eigenvalues of A and A_j be arranged in increasing order. For each $i=1,2,\ldots,n-1$ we have

$$\max_{1\le j\le n}\lambda_i(A_j)\ge\frac{n-i}{n}\lambda_1(A)+\frac{i}{n}\lambda_{i+1}(A)$$

$$\min_{1\le j\le n}\lambda_i(A_j)\le\frac{n-i}{n}\lambda_i(A)+\frac{i}{n}\lambda_n(A)$$

and

$$\max_{1\le j\le n}\lambda_{n-1}(A_j)-\min_{1\le j\le n}\lambda_1(A_j)\ge\left(\frac{n-2}{n}\right)^{1/2}[\lambda_n(A)-\lambda_1(A)]$$

If all the eigenvalues of A are nonnegative, that is, if A is positive semidefinite, the first of these three inequalities implies that there is at least one principal submatrix A_j for which

$$\lambda_{n-1}(A_j)\ge\frac{n-1}{n}\lambda_n(A)$$

Thus, it is not possible for the spectral radius of *every* principal submatrix of a positive semidefinite Hermitian matrix to be "small."

One may wish to delete several rows and the corresponding columns from a Hermitian matrix. The remaining matrix is a principal submatrix of the original matrix. The following result can be obtained by repeated application of the interlacing inequalities (4.3.9), but it is just as easy to prove the assertions directly from the Courant-Fischer theorem. This result is sometimes called the *inclusion principle.*

4.3.15 Theorem. Let $A\in M_n$ be a Hermitian matrix, let r be an integer with $1\le r\le n$, and let A_r denote any r-by-r principal submatrix of A (obtained by deleting $n-r$ rows and the corresponding columns from A). For each integer k such that $1\le k\le r$ we have

$$\lambda_k(A)\le\lambda_k(A_r)\le\lambda_{k+n-r}(A)$$

Proof: Suppose $A_r \in M_r$ is formed by deleting rows $i_1, \ldots, i_{n-r}$ and the corresponding columns from A, and let $1 \le k \le r$. Use (4.2.12) to write

$$\lambda_{k+n-r}(A) = \min_{w_1,\ldots,w_{r-k} \in C^n} \max_{\substack{x \ne 0,\, x \in C^n \\ x \perp w_1,\ldots,w_{r-k}}} \frac{x^*Ax}{x^*x}$$

$$\ge \min_{w_1,\ldots,w_{r-k} \in C^n} \max_{\substack{x \ne 0,\, x \in C^n \\ x \perp w_1,\ldots,w_{r-k} \\ x \perp e_{i_1},\ldots,e_{i_{n-r}}}} \frac{x^*Ax}{x^*x}$$

$$= \min_{v_1,\ldots,v_{r-k} \in C^r} \max_{\substack{y \ne 0,\, y \in C^r \\ y \perp v_1,\ldots,v_{r-k}}} \frac{y^*A_ry}{y^*y} = \lambda_k(A_r)$$

Again, assuming $1 \le k \le r$, use (4.2.13) to write

$$\lambda_k(A) = \max_{w_1,\ldots,w_{k-1} \in C^n} \min_{\substack{x \ne 0,\, x \in C^n \\ x \perp w_1,\ldots,w_{k-1}}} \frac{x^*Ax}{x^*x}$$

$$\le \max_{w_1,\ldots,w_{k-1} \in C^n} \min_{\substack{x \ne 0,\, x \in C^n \\ x \perp w_1,\ldots,w_{k-1} \\ x \perp e_{i_1},\ldots,e_{i_{n-r}}}} \frac{x^*Ax}{x^*x}$$

$$= \max_{v_1,\ldots,v_{k-1} \in C^r} \min_{\substack{y \ne 0,\, y \in C^r \\ y \perp v_1,\ldots,v_{k-1}}} \frac{y^*A_ry}{y^*y} = \lambda_k(A_r)$$ □

The following easy consequence of Theorem (4.3.15) is known as the *Poincaré separation theorem.* It can be employed in situations (such as in quantum mechanics) in which one has information about the inner products $u_i^*Au_j$.

4.3.16 **Corollary.** Let $A \in M_n$ be Hermitian, let r be a given integer with $1 \le r \le n$, and let $u_1, \ldots, u_r \in \mathbf{C}^n$ be r given orthonormal vectors. Let $B_r \equiv [u_i^*Au_j] \in M_r$. If the eigenvalues of A and B_r are arranged in increasing order (4.2.1), we have

$$\lambda_k(A) \le \lambda_k(B_r) \le \lambda_{k+n-r}(A), \qquad k = 1, 2, \ldots, r \tag{4.3.17}$$

Proof: If $r < n$, choose $n-r$ additional vectors $u_{r+1}, \ldots, u_n$ so that $\{u_1, \ldots, u_r, u_{r+1}, \ldots, u_n\}$ is an orthonormal set, and let $U = [u_1 \ldots u_n] \in M_n$. The matrix U is unitary, U^*AU has the same eigenvalues as A, and the given matrix B_r is a principal submatrix of U^*AU obtained by deleting the last $n-r$ rows and columns. The assertion now follows from (4.3.15). □

The matrix $B_r \in M_r$ in the preceding result can be written as $B_r = U^*AU$, where $U \in M_{n,r}$ is a matrix with r orthonormal columns. Since $\operatorname{tr} B_r = \lambda_1(B_r) + \cdots + \lambda_r(B_r)$, the following extremal characterization follows from summing the inequalities (4.3.17).

4.3.18 **Corollary.** Let $A \in M_n$ be Hermitian and let r be a given integer with $1 \le r \le n$. Then

$$\lambda_1(A) + \cdots + \lambda_r(A) = \min_{U^*U = I \in M_r} \operatorname{tr} U^*AU \tag{4.3.19}$$

$$\lambda_{n-r+1}(A) + \cdots + \lambda_n(A) = \max_{U^*U = I \in M_r} \operatorname{tr} U^*AU \tag{4.3.20}$$

$U \in M_{n,r}$

Equality holds in (4.3.19) if the columns of U are chosen to be orthonormal eigenvectors corresponding to the r smallest eigenvalues of A. A similar choice yields equality in (4.3.20). These two inequalities may be thought of as generalizations of the Rayleigh–Ritz theorem (4.2.2). They can be used to prove many other interesting inequalities.

Sometimes one has bounds on the behavior of the quadratic form x^*Ax on a subspace. The Courant–Fischer theorem can be employed in this case to give bounds on the eigenvalues of A.

4.3.21 **Theorem.** Let $A \in M_n$ be Hermitian, let k be a given integer with $1 \le k \le n$, let the eigenvalues of A be arranged in increasing order (4.2.1), and let S_k be a given k-dimensional subspace of $\mathbf{C}^n$. If there exists a constant c_2 such that $x^*Ax \ge c_2x^*x$ for all $x \in S_k$, then $\lambda_n \ge \lambda_{n-1} \ge \cdots \ge \lambda_{n-k+1} \ge c_2$. If there exists a constant c_1 such that $x^*Ax \le c_1x^*x$ for all $x \in S_k$, then $c_1 \ge \lambda_k \ge \cdots \ge \lambda_1$.

Proof: Let $u_1, \ldots, u_{n-k}$ be $n-k$ orthonormal vectors that span $S_k^\perp$. Use (4.2.13) to write

$$c_2 \le \min_{\substack{x \ne 0 \\ x \in S_k}} \frac{x^*Ax}{x^*x} = \min_{\substack{x \ne 0 \\ x \perp u_1,\ldots,u_{n-k}}} \frac{x^*Ax}{x^*x} \le \max_{w_1,\ldots,w_{n-k} \in C^n} \min_{\substack{x \ne 0 \\ x \perp w_1,\ldots,w_{n-k}}} \frac{x^*Ax}{x^*x} = \lambda_{n-k+1} \tag{4.3.22}$$

Similarly, we can use (4.2.12) to write

$$c_1 \ge \max_{\substack{x\neq 0\\ x\in S_k}} \frac{x^*Ax}{x^*x} = \max_{\substack{x\neq 0\\ x\perp u_1,\ldots,u_{n-k}}} \frac{x^*Ax}{x^*x}$$

$$\ge \min_{w_1,\ldots,w_{n-k}\in \mathbf{C}^n} \max_{\substack{x\neq 0\\ x\perp w_1,\ldots,w_{n-k}}} \frac{x^*Ax}{x^*x} = \lambda_k$$

□

4.3.23 **Corollary.** If $A\in M_n$ is Hermitian and if $x^*Ax\ge 0$ for all vectors x in a k-dimensional subspace, then A has at least k nonnegative eigenvalues. If $x^*Ax>0$ for all nonzero vectors x in a k-dimensional subspace, then A has at least k positive eigenvalues.

Proof: The first assertion follows from the preceding theorem with $c_2=0$. If $\lambda_{n-k+1}=0$, then the inequalities (4.3.22) show that

$$0=\min_{\substack{x\neq 0\\ x\in S_k}} \frac{x^*Ax}{x^*x} = \min_{\substack{x^*x=1\\ x\in S_k}} x^*Ax$$

But S_k is finite-dimensional, so the set $D=\{x\in S_k: x^*x=1\}$ is a compact set [see (5.5.6)] and the continuous function x^*Ax achieves its minimum on D for some $x_0\in S_k$ such that $x^*x=1$; in particular, $x_0\neq 0$. But then $x_0^*Ax_0=0$ contradicts the assumption that $x^*Ax>0$ whenever $x\in S_k$ and $x\neq 0$. □

Both the eigenvalues and the main diagonal elements of a Hermitian matrix are real numbers, and the sum of the eigenvalues is the same as the sum of the main diagonal elements (the trace). The precise relationship between the main diagonal elements and the eigenvalues is given by the notion of *majorization*.

4.3.24 **Definition.** Let $\alpha=[\alpha_i]\in \mathbf{R}^n$ and $\beta=[\beta_i]\in\mathbf{R}^n$ be given. The vector β is said to *majorize* the vector α if

$$\min\left\{\sum_{j=1}^{k}\beta_{i_j}: 1\le i_1<\cdots<i_k\le n\right\}\ge \min\left\{\sum_{j=1}^{k}\alpha_{i_j}: 1\le i_1<\cdots<i_k\le n\right\}$$

for all $k=1,2,\ldots,n$ with *equality* for $k=n$. If we arrange the entries of α and β in increasing order $\alpha_{j_1}\le\alpha_{j_2}\le\cdots\le\alpha_{j_n}$, $\beta_{m_1}\le\beta_{m_2}\le\cdots\le\beta_{m_n}$, the defining inequalities can be restated in the equivalent form

$$\sum_{i=1}^{k}\beta_{m_i}\ge\sum_{i=1}^{k}\alpha_{j_i}\quad\text{for all}\quad k=1,2,\ldots,n \tag{4.3.25}$$

with *equality* for $k=n$.

Thus, the real vector β majorizes the real vector α if the sum of the k smallest entries of β is greater than or equal to the sum of the k smallest entries of α for $k=1,2,\ldots,n-1$ and the sums of the entries of β and α are equal. Notice that the entries of β and α may be permuted arbitrarily without affecting whether β majorizes α.

The notion of majorization is an important one that arises in many places in matrix theory as the precise relationship between two sets of real numbers. One example of this phenomenon is the following theorem of Schur (1923).

4.3.26 **Theorem.** Let $A\in M_n$ be Hermitian. The vector of diagonal entries of A majorizes the vector of eigenvalues of A.

Proof: The proof is by induction on the dimension. For $n=1$, there is nothing to show, so we suppose that the result is valid for Hermitian matrices of dimension k for all $k\le n-1$. Let $A\equiv[a_{ij}]\in M_n$ be a given Hermitian matrix and let $A_1\in M_{n-1}$ be the principal submatrix of A obtained by deleting the row and column corresponding to the *largest* diagonal entry of A. Let $\lambda_1\le\cdots\le\lambda_n$ be the ordered eigenvalues of A, let $\lambda_1'\le\cdots\le\lambda_{n-1}'$ be the ordered eigenvalues of A_1, and let $a_{i_1i_1}\le a_{i_2i_2}\le\cdots\le a_{i_ni_n}$ be a rearrangement of the diagonal entries into increasing order. By the induction hypothesis, we have

$$\sum_{j=1}^{k} a_{i_ji_j}\ge\sum_{j=1}^{k}\lambda_j'\quad\text{for all}\quad k=1,,\ldots,n-1$$

Because of interlacing [Theorem (4.3.8)], we also have

$$\lambda_1\le\lambda_1'\le\lambda_2\le\lambda_2'\le\cdots\le\lambda_{n-1}'\le\lambda_n$$

and hence

$$\sum_{j=1}^{k}\lambda_j'\ge\sum_{j=1}^{k}\lambda_j\quad\text{for all}\quad k=1,\ldots,n-1$$

Thus,

$$\sum_{j=1}^{k} a_{i_ji_j}\ge\sum_{j=1}^{k}\lambda_j,\qquad k=1,\ldots,n-1$$

and equality holds for $k=n$ because the trace is the sum of the eigenvalues. □

Majorization is also useful in expressing the relationship between the eigenvalues of a sum and those of its summands.

4.3.27 Theorem. Let $A, B \in M_n$ be Hermitian matrices, and let $\lambda(A) = [\lambda_i(A)]$, $\lambda(B) = [\lambda_i(B)]$, and $\lambda(A+B) = [\lambda_i(A+B)]$ denote the column vectors in $\mathbf{R}^n$ whose components are the eigenvalues of A, B, and $A+B$ arranged in increasing order (4.2.1). The vector $\lambda(A+B)$ majorizes the vector $\lambda(A)+\lambda(B)$.

Proof: For any $k = 1, 2, \ldots, n$, use (4.3.18) to write

$$\begin{aligned}\sum_{i=1}^{k} \lambda_i(A+B) &= \min_{U^*U=I\in M_k} \operatorname{tr} U^*(A+B)U \\ &= \min_{U^*U=I\in M_k} (\operatorname{tr} U^*AU + \operatorname{tr} U^*BU) \\ &\geq \min_{U^*U=I\in M_k} \operatorname{tr} U^*AU + \min_{U^*U=I\in M_k} \operatorname{tr} U^*BU \\ &= \sum_{i=1}^{k} \lambda_i(A) + \sum_{i=1}^{k} \lambda_i(B) = \sum_{i=1}^{k} (\lambda_i(A)+\lambda_i(B))\end{aligned}$$

Since $\operatorname{tr}(A+B) = \operatorname{tr} A + \operatorname{tr} B$, we have equality for $k = n$. □

We have alluded to the fact that majorization is the *precise* relationship between the main diagonal elements of a Hermitian matrix and its eigenvalues, but we have established only half of this relationship in (4.3.26). To show the other half, we need the following technical lemma.

4.3.28 Lemma. Let $n \geq 2$ and let $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_n$ and $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_n$ be given real numbers. If the vector $\beta = [\beta_i]$ majorizes the vector $\alpha = [\alpha_i]$, then there are $n-1$ real numbers $\gamma_1, \ldots, \gamma_{n-1}$ such that

$$\alpha_1 \leq \gamma_1 \leq \alpha_2 \leq \gamma_2 \leq \alpha_3 \leq \cdots \leq \alpha_{n-1} \leq \gamma_{n-1} \leq \alpha_n$$

and such that $\beta' = [\beta_1, \ldots, \beta_{n-1}]^T \in \mathbf{R}^{n-1}$ majorizes $\gamma = [\gamma_1, \ldots, \gamma_{n-1}]^T \in \mathbf{R}^{n-1}$.

Proof: If $n = 2$, we have $\alpha_1 \leq \beta_1$ and $\alpha_1 + \alpha_2 = \beta_1 + \beta_2$, or

$$\alpha_2 = (\beta_1 - \alpha_1) + \beta_2 \geq \beta_2 \geq \beta_1$$

Thus, $\alpha_1 \leq \beta_1 \leq \alpha_2$, so we can choose $\gamma_1 = \beta_1$ and satisfy the stated conditions. Now assume that $n \geq 2$ and let $\Delta = \{[\delta_1, \ldots, \delta_{n-1}]^T\} \subset \mathbf{R}^{n-1}$ denote the set of points defined by the inequalities

$$\alpha_1 \leq \delta_1 \leq \alpha_2 \leq \delta_2 \leq \alpha_3 \leq \cdots \leq \alpha_{n-1} \leq \delta_{n-1} \leq \alpha_n \tag{4.3.29a}$$

$$\sum_{i=1}^{k} \delta_i \leq \sum_{i=1}^{k} \beta_i, \qquad k = 1, 2, \ldots, n-2 \tag{4.3.29b}$$

Because β majorizes α, the point $\delta = \hat{\alpha} = [\alpha_1, \ldots, \alpha_{n-1}]^T$ is always in Δ, so the set Δ is always nonempty. The set Δ is evidently bounded and closed, hence compact, and it is easily seen to be convex. If $\delta = [\delta_1, \ldots, \delta_{n-1}]^T \in \Delta$, define $f(\delta) \equiv \delta_1 + \delta_2 + \cdots + \delta_{n-1}$. Notice that $f(\hat{\alpha}) = \alpha_1 + \cdots + \alpha_{n-1} \leq \beta_1 + \cdots + \beta_{n-1}$. If we can show that there is some $\hat{\delta} \in \Delta$ with $f(\hat{\delta}) \geq \beta_1 + \cdots + \beta_{n-1}$, then by convexity of Δ we will have $t\hat{\alpha} + (1-t)\hat{\delta} \in \Delta$ for all $t \in [0, 1]$ and $g(t) \equiv f(t\hat{\alpha} + (1-t)\hat{\delta})$ will be a continuous function with

$$g(0) \geq \beta_1 + \cdots + \beta_{n-1} \geq g(1)$$

From this we can conclude that for some $t_0 \in [0, 1]$, we have $g(t_0) = \beta_1 + \cdots + \beta_{n-1}$. The point $\gamma = [\gamma_i] = t_0\hat{\alpha} + (1-t_0)\hat{\delta}$ will satisfy the stated conditions.

Since $f(\cdot)$ is a continuous function on the compact set Δ, there is a point $\hat{\delta} \in \Delta$ such that

$$\max_{\delta\in\Delta} f(\delta) = f(\hat{\delta}) \tag{4.3.30}$$

We shall show that $f(\hat{\delta}) \geq \beta_1 + \cdots + \beta_{n-1}$. A maximizing point $\hat{\delta} \in \Delta$ satisfies the inequalities (4.3.29a and b), and hence

$$\hat{\delta}_k \leq \alpha_{k+1}, \qquad k = 1, 2, \ldots, n-1 \tag{4.3.31a}$$

$$\sum_{i=1}^{k} \hat{\delta}_i \leq \sum_{i=1}^{k} \beta_i, \qquad k = 1, 2, \ldots, n-2 \tag{4.3.31b}$$

If *all* the inequalities (4.3.31b) are strict, and if at least one of the inequalities (4.3.31a) is not an equality, then at least one component of $\hat{\delta}$ can be increased with a consequent increase in the value of $f(\hat{\delta})$. Since this contradicts the extremal property (4.3.30), we conclude that all the inequalities (4.3.31a) must be equalities, $\hat{\delta} = [\alpha_2, \alpha_3, \ldots, \alpha_n]^T$, and $f(\hat{\delta}) = \alpha_2 + \cdots + \alpha_n = (\alpha_1 + \alpha_2 + \cdots + \alpha_n) - \alpha_1 = (\beta_1 + \beta_2 + \cdots + \beta_n) - \alpha_1 = (\beta_1 + \cdots + \beta_{n-1}) + \beta_n - \alpha_1 \geq (\beta_1 + \cdots + \beta_{n-1}) + \beta_1 - \alpha_1 \geq \beta_1 + \cdots + \beta_{n-1}$, which is what we wish to show.

If *not all* the inequalities (4.3.31b) are strict, then equality holds for at least one value of k. Let r denote the largest such value of k. Then

$$\sum_{i=1}^{r} \hat{\delta}_i = \sum_{i=1}^{r} \beta_i$$

$$\sum_{i=1}^{k} \hat{\delta}_i < \sum_{i=1}^{k} \beta_i \quad \text{for} \quad k = r+1, \ldots, n-2$$

By the same argument as in the preceding paragraph, we must have $\hat{\delta}_k = \alpha_{k+1}$ for $k = r+1, \ldots, n-1$. Thus,

$$\begin{aligned} f(\hat{\delta}) &= (\hat{\delta}_1 + \cdots + \hat{\delta}_r) + (\hat{\delta}_{r+1} + \cdots + \hat{\delta}_{n-1}) \\ &= (\beta_1 + \cdots + \beta_r) + (\alpha_{r+2} + \cdots + \alpha_n) \\ &= (\beta_1 + \cdots + \beta_{n-1}) + (\alpha_1 + \cdots + \alpha_n) \\ &\quad - (\alpha_1 + \cdots + \alpha_{r+1}) - (\beta_{r+1} + \cdots + \beta_{n-1}) \\ &= (\beta_1 + \cdots + \beta_{n-1}) + (\beta_1 + \cdots + \beta_n) \\ &\quad - (\alpha_1 + \cdots + \alpha_{r+1}) - (\beta_{r+1} + \cdots + \beta_{n-1}) \\ &= (\beta_1 + \cdots + \beta_{n-1}) + [(\beta_1 + \cdots + \beta_{r+1}) - (\alpha_1 + \cdots + \alpha_{r+1})] \\ &\quad + (\beta_{r+2} + \cdots + \beta_n) - (\beta_{r+1} + \cdots + \beta_{n-1}) \\ &\geq (\beta_1 + \cdots + \beta_{n-1}) + (\beta_{r+2} - \beta_{r+1}) \\ &\quad + (\beta_{r+3} - \beta_{r+2}) + \cdots + (\beta_n - \beta_{n-1}) \\ &\geq \beta_1 + \cdots + \beta_{n-1} \end{aligned}$$

□

We can now prove a converse to (4.3.26).

4.3.32 **Theorem.** Let $n \geq 1$ and let $a_1 \leq a_2 \leq \cdots \leq a_n$ and $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ be given real numbers. If the vector $a = [a_i]$ majorizes the vector $\lambda = [\lambda_i]$, then there exists a real symmetric matrix $A = [a_{ij}] \in M_n(\mathbf{R})$ such that $a_{ii} = a_i$ for $i = 1, 2, \ldots, n$ and such that $\{\lambda_i\}$ is the set of eigenvalues of A.

Proof: The assertion is trivial for $n = 1$. Suppose it has been proved for all such vectors a and λ with at most $n-1$ elements. By the lemma there exist real numbers $\gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_{n-1}$ such that

$$\lambda_1 \leq \gamma_1 \leq \lambda_2 \leq \cdots \leq \lambda_{n-1} \leq \gamma_{n-1} \leq \lambda_n$$

and such that $a' = [a_1, \ldots, a_{n-1}]^T$ majorizes $\gamma = [\gamma_i]^T \in \mathbf{R}^{n-1}$. By the induction hypothesis there is a real symmetric matrix $B = [b_{ij}] \in M_{n-1}$ with $b_{ii} = a_i$ for $i = 1, 2, \ldots, n-1$ such that $\{\gamma_i\}$ is the set of eigenvalues of B. If $\Gamma = \operatorname{diag}(\gamma_1, \gamma_2, \ldots, \gamma_{n-1}) \in M_{n-1}(\mathbf{R})$, there is a real orthogonal matrix $Q \in M_{n-1}(\mathbf{R})$ such that $B = Q\Gamma Q^T$. By Theorem (4.3.10) there is a real symmetric matrix

$$\hat{A} \equiv \begin{bmatrix} \Gamma & y \\ y^T & \alpha \end{bmatrix} \in M_n(\mathbf{R}), \qquad y \in \mathbf{R}^{n-1}, \qquad \alpha \in \mathbf{R}$$

that has eigenvalues $\{\lambda_i\}$. If we set

$$A \equiv \begin{bmatrix} Q & 0 \\ 0 & 1 \end{bmatrix} \hat{A} \begin{bmatrix} Q^T & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} Q\Gamma Q^T & Qy \\ (Qy)^T & \alpha \end{bmatrix} = \begin{bmatrix} B & Qy \\ (Qy)^T & \alpha \end{bmatrix}$$

then A has eigenvalues $\{\lambda_i\}$ and has main diagonal entries $a_1, a_2, \ldots, a_{n-1}, \alpha$. But $\operatorname{tr} A = a_1 + \cdots + a_{n-1} + \alpha = \lambda_1 + \cdots + \lambda_n = a_1 + \cdots + a_n$ by majorization, so $\alpha = a_n$ and A has the correct diagonal entries. □

The preceding result not only completes the circle of implications dealing with the relationships between the main diagonal entries and eigenvalues of a Hermitian matrix, but also permits us to clarify the geometric meaning of the majorization relation itself. A *doubly stochastic* matrix $A \in M_n$ has n^2 nonnegative entries such that the sum of the entries in every row and every column is +1. Birkhoff's theorem (8.7.1) guarantees that every doubly stochastic matrix is a convex combination of finitely many permutation matrices and conversely.

4.3.33 **Theorem.** Let $\alpha = [\alpha_i] \in \mathbf{R}^n$ and $\beta = [\beta_i] \in \mathbf{R}^n$ be two given real vectors. The following are equivalent:

(a) β majorizes α;
(b) There is a doubly stochastic matrix $S \in M_n$ such that $\beta = S\alpha$; and
(c) $\beta \in \left\{ \sum_{i=1}^{N} p_i \alpha_{\pi_i} \right\}$, where $1 \leq N < \infty$, $p_i \geq 0$, $\sum_{i=1}^{N} p_i = 1$ and $\alpha_{\pi_i} \in \mathbf{R}^n$ is a vector whose components are some permutation of the components of the given vector α.

Proof: If we assume (a), then by (4.3.32) there is a real symmetric matrix $B = [b_{ij}] \in M_n$ with main diagonal $b_{ii} = \beta_i$ and eigenvalues $\lambda_i(B) = \alpha_i$. By the spectral theorem there is a unitary (even real orthogonal) matrix $U = [u_{ij}] \in M_n$ such that $B = U\Lambda U^*$, where $\Lambda = \operatorname{diag}(\alpha_1, \ldots, \alpha_n)$, and consideration of the main diagonal entries of B shows that $\beta = S\alpha$, where $S = [s_{ij}] \in M_n$ is given by $s_{ij} = |u_{ij}|^2$. Such a matrix S has every row and column sum equal to 1 since every row and column of U is a unit vector, so S is a doubly stochastic matrix (of a special type known as an *orthostochastic* matrix). This shows that (a) implies (b).

A proof that (b) implies (a) is outlined in Problem 9 at the end of this section.

If we assume (b), then Birkhoff's theorem (8.7.1) shows that

$$S = \sum_{i=1}^{N} p_i P_i, \qquad \text{where } p_i \geq 0, \qquad \sum_{i=1}^{N} p_i = 1$$

and each P_i is a permutation matrix. Thus,

$$\beta = S\alpha = \sum_{i=1}^{N} p_i P_i \alpha = \sum_{i=1}^{N} p_i \alpha_{\pi_i}, \qquad \text{where } P_i \alpha \equiv \alpha_{\pi_i}$$

This identity also establishes the reverse implication. □

Thus, the collection of all vectors $\beta = [\beta_1, \ldots, \beta_n]^T$ that majorize a given vector $\alpha = [\alpha_1, \ldots, \alpha_n]^T$ may be obtained by computing the $n!$ vectors (not all distinct if not all the α_i terms are distinct) obtained by permuting the n components of the vector α and then forming the convex hull of these vectors.

Remark: While there is universal agreement that the basic notion of majorization is important, there is not universal agreement on notation. Some authors define majorization with the inequality reversed in (4.3.25), and some define it with respect to decreasing arrangements of the two sets. For this reason, one should use caution when using or quoting results about majorization from different sources. See Problem 11 for a justification of our choice of the definition of majorization.

Problems

1. Recall that the spectral radius of a matrix $A \in M_n$ is the quantity

$$\rho(A) = \max\{|\lambda_i(A)|\}$$

Let $A, B \in M_n$ be Hermitian. Use Weyl's theorem (4.3.1) to show that

$$\lambda_1(B) \le \lambda_k(A+B) - \lambda_k(A) \le \lambda_n(B)$$

and hence that

$$|\lambda_k(A+B) - \lambda_k(A)| \le \rho(B)$$

for all $k = 1, 2, \ldots, n$. This is a simple example of a *perturbation theorem* for the eigenvalues of a Hermitian matrix [cf. (6.3)].

2. Show how to use only the first set of inequalities derived in the proof of Theorem (4.3.4) to obtain all the inequalities asserted in the theorem. *Hint:* $A = (A \pm zz^*) \mp zz^*$.

3. Provide the details to show that (4.3.6b) is equivalent to (4.3.6a).

4. The only case of Theorem (4.3.6) used in the proof of Weyl's theorem (4.3.7) is $\lambda_n(A+B) \ge \lambda_{n-r}(A)$, where B has rank at most r. Show that this case can be proved without using the Courant–Fischer theorem by providing the details for the following argument. Suppose $B = \beta_1 y_1 y_1^* + \cdots + \beta_r y_r y_r^*$ and let $A = U\Lambda U^*$ with $U = [u_1 \cdots u_n]$ unitary. Then there exist $r+1$ scalars $\alpha_{n-r}, \alpha_{n-r+1}, \ldots, \alpha_n$ such that the vector $x \equiv \alpha_{n-r} u_{n-r} + \cdots + \alpha_n u_n$ satisfies $x \perp y_i$ for all $i = 1, 2, \ldots, r$ and $x^*x = |\alpha_{n-r}|^2 + \cdots + |\alpha_n|^2 = 1$. Then

$$\lambda_n(A-B) \ge x^*(A-B)x = \sum_{i=n-r}^{n} |\alpha_i|^2 \lambda_i(A) \ge \lambda_{n-r}(A)$$

5. Deduce Theorem (4.3.15) by applying Theorem (4.3.8) $n-r$ times.

6. Show that Weyl's simple inequalities (4.3.2) need not hold if A and B are not Hermitian. *Hint:* Consider $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$.

7. If $A, B \in M_n$ are Hermitian matrices with eigenvalues arranged in increasing order, and if $1 \le k \le n$, show that

$$\lambda_k(A+B) \le \min\{\lambda_i(A) + \lambda_j(B) : i + j = k + n\}$$

8. Provide the details for the case of coincident λ_i terms in the proof of Theorem (4.3.10). *Hint:* If λ_1 is a k-fold root of $g(t) = 0$ with $k \ge 2$, show that one obtains (4.3.14) in which the factor $(t-\lambda_1)^{k-1}$ in $g'(t)$ in the denominator of (4.3.14) is canceled into a factor $(t-\lambda_1)^{k-1}$ in $f(t)$ in the numerator of (4.3.14).

9. Let $S = [s_{ij}] \in M_n$ be a doubly stochastic matrix (8.7), and let $x \in \mathbf{R}^n$ be a real vector. Show that Sx majorizes x. *Hint:* Let $y = Sx$. It suffices to assume that $y_1 \le \cdots \le y_n$ and $x_1 \le \cdots \le x_n$, for if not, one could consider Py and Qx for suitable permutation matrices P and Q; PSQ^T is still doubly stochastic. Let $w_j^{(k)} = \sum_{i=1}^k s_{ij}$, so $0 \le w_j^{(k)} \le 1$ and $\sum_{j=1}^n w_j^{(k)} = k$. Show that

$$\sum_{i=1}^{k} (y_i - x_i) = \sum_{j=1}^{n} w_j^{(k)} x_j - \sum_{i=1}^{k} x_i + x_k \left(k - \sum_{j=1}^{n} w_j^{(k)}\right)$$

$$= \sum_{j=1}^{k} (1 - w_j^{(k)})(x_k - x_j) + \sum_{j=k+1}^{n} w_j^{(k)}(x_j - x_k)$$

and that all the terms in the latter sums are nonnegative.

10. Give another proof of (4.3.26) using the following ideas: If $A = [a_{ij}] \in M_n$ is Hermitian, then $A = U\Lambda U^*$ with $U = [u_{ij}] \in M_n$ unitary and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ a real diagonal matrix. Let $a = [a_{11}, a_{22}, \ldots, a_{nn}]^T$ be the vector consisting of the main diagonal entries of A and let $x = [\lambda_1, \lambda_2, \ldots, \lambda_n]^T$. Show that $a = Px$, where $P = [p_{ij}] = [|u_{ij}|^2]$. Show that P is doubly stochastic and use Problem 9.

11. Let $x = [x_1, \ldots, x_n]^T$ and $y = [y_1, \ldots, y_n]^T$ be two given nonnegative real vectors and suppose that y majorizes x. Show that $y_1 \cdots y_n \ge x_1 \cdots x_n$. *Hint:* Use (4.3.32) to construct a real symmetric matrix $A = [a_{ij}] \in M_n(\mathbf{R})$ with main diagonal entries $a_{ii} \equiv y_i$ and eigenvalues $\lambda_i(A) = x_i$. Then Hadamard's inequality (7.8.1) says that $a_{11} \cdots a_{nn} \ge \det A = \lambda_1 \cdots \lambda_n$. *Remark:* It is this result that motivates our choice of the definition (4.3.24) of majorization. If one chooses the opposite direction for the inequality in (4.3.24), then the consequence is that the direction of

the inequality in Problem 11 is reversed; that is, if y "majorizes" x in this sense, then the product of the y_i is *less* than the product of the x_i. We prefer a definition in which the inequality for the products goes in the same direction as the majorization.

12. Let $A \in M_n$ be a Hermitian matrix with positive eigenvalues $0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ and let $1 \le r \le n$ be a given integer. Use Problem 11 to prove that

$$\lambda_1 \lambda_2 \cdots \lambda_r = \min (u_1^* A u_1)(u_2^* A u_2) \cdots (u_r^* A u_r)$$

where the minimum is over all sets of orthonormal vectors $\{u_1, u_2, \ldots, u_r\} \subset \mathbf{C}^n$. Explain why this result may be thought of as a multiplicative analog of (4.3.19), as a generalization of Hadamard's inequality (7.8.1), and as a chain of inequalities that connects the Rayleigh–Ritz theorem (4.2.2) to Hadamard's inequality. *Hint:* The case $r = n$ is (7.8.1). What is the case $r = 1$? If $2 \le r \le n$, use (4.3.19) to show that the vector $[u_1^* A u_1, u_2^* A u_2, \ldots, u_r^* A u_r]^T$ majorizes the vector $[\mu_1, \ldots, \mu_r]^T$, where $\mu_i = \lambda_i$ for $i = 1, 2, \ldots, r-1$ and

$$\mu_r = (u_1^* A u_1 + \cdots + u_{r-1}^* A u_{r-1}) - (\lambda_1 + \cdots + \lambda_{r-1}) + u_r^* A u_r \ge u_r^* A u_r$$

Now use Problem 11 to show that $\lambda_1 \cdots \lambda_{r-1} u_r^* A u_r \le \prod_{i=1}^r u_i^* A u_i$.

13. Let $A = [a_{ij}] \in M_n$ be a Hermitian matrix with nonnegative eigenvalues $0 \le \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$. Show for each $r = 1, 2, \ldots, n$ that the product $\lambda_1 \cdots \lambda_r$ is less than or equal to the product of the r smallest main diagonal entries of A.

14. If $A, B \in M_n$ are Hermitian and $A - B$ has only nonnegative eigenvalues, show that $\lambda_i(A) \ge \lambda_i(B)$ for all $i = 1, 2, \ldots, n$.

15. Use (4.3.18) to prove (4.3.26). *Hint:* By a permutation, arrange $A = [a_{ij}]$ so that $a_{11} \le a_{22} \le \cdots \le a_{nn}$. Then take $U = [e_1\ e_2\ \cdots\ e_r] \in M_{n,r}$ and show that $\lambda_1(A) + \cdots + \lambda_r(A) \le \operatorname{tr} U^* A U = a_{11} + \cdots + a_{rr}$.

16. Suppose $A \in M_n$ is Hermitian. Let $\lambda_1 \le \cdots \le \lambda_n$ be the eigenvalues of A and $\lambda_{i,1} \le \cdots \le \lambda_{i,n-1}$ be the eigenvalues of the $(n-1)$-by-$(n-1)$ principal submatrix $A(\{i\}')$. Show that

$$\lambda_1 \le \lambda_{i,1} \le \lambda_2 \le \lambda_{i,2} \le \cdots \le \lambda_{i,n-1} \le \lambda_n$$

These *interlacing* inequalities are often attributed to Cauchy. Show also that these inequalities imply the inequalities in (4.3.15). *Hint:* Use (4.3.8) or (4.3.15).

17. If $A = [a_{ij}] \in M_n$ is Hermitian, and if $a_{ii} = \lambda_n$ for some i, show that $a_{ik} = a_{ki} = 0$ for all $k = 1, 2, \ldots, n$, $k \ne i$, and similarly if $a_{ii} = \lambda_1$. *Hint:* Show this by explicit calculation for $n = 2$ and apply interlacing.

18. Let $A \in M_n$ be Hermitian and let $a_i = \det A(\{1, 2, \ldots, i\})$, $i = 1, 2, \ldots, n$. If all $a_i \ne 0$, show that the number of negative eigenvalues of A is equal to the number of sign changes in the sequence $+1, a_1, a_2, \ldots, a_n$. In particular, if all these principal minors are positive, A has no negative eigenvalues at all. What happens if some $a_i = 0$? *Hint:* Use interlacing.

19. Let $A = [a_{ij}] \in M_n$ be normal. Show that if A has a "small" column or row, then it must also have a "small" eigenvalue. More precisely, let the squares of the absolute values of the eigenvalues of A $\{|\lambda_i|^2 : i = 1, \ldots, n\}$ be placed in nondecreasing order and denote the resulting ordered values by $v_1^2 \le v_2^2 \le v_3^2 \le \cdots \le v_n^2$. Let the square roots of the sums of squares of the absolute values of the entries in the rows $\{(\sum_{k=1}^n |a_{ik}|^2)^{1/2} : i = 1, \ldots, n\}$ (or the columns) be placed in nondecreasing order and denote the resulting ordered values by $R_1 \le R_2 \le \cdots \le R_n$. Show that

$$\sum_{i=1}^k v_i^2 \le \sum_{i=1}^k R_i^2 \quad \text{for} \quad k = 1, \ldots, n$$

with a similar upper bound involving the column sums of squares. *Hint:* The quantities v_i^2 are the eigenvalues of the Hermitian matrix AA^*. What are the main diagonal entries of AA^*? Use majorization and Theorem (4.3.26). Consider A^*A for the column sum inequalities.

Further Readings. For more information about majorization, see [MOl]. For a discussion of the general interlacing inequalities among the eigenvalues of principal submatrices [mentioned following Theorem (4.3.10)] see C. R. Johnson and H. A. Robinson, "Eigenvalue Inequalities for Principal Submatrices," *Lin. Alg. Appl.* 37 (1981), 11–22. The argument outlined in Problem 4 is the original proof from H. Weyl, "Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung," *Math. Annalen* 71 (1912), 441ff.; see the proof of the lemma on pp. 444–445. Weyl states and proves his result in terms of integral equations, but the translation to linear algebraic form is immediate.

4.4 Complex symmetric matrices

A matrix $A \in M_n$ is *symmetric* if $A = A^T$. In many instances, symmetric matrices under study also have only real entries, so they are real Hermitian matrices and all the results discussed so far in this chapter apply to them.

There are some circumstances in which one confronts complex symmetric matrices, however. One example is in the study of regular analytic

mappings of the unit disc in the complex plane. If $f(z)$ is a regular analytic function on the unit disc, and if $f(z)$ is normalized so that $f(0)=0$ and $f'(0)=1$, then $f(z)$ is one-to-one (sometimes called *univalent* or *schlicht*) if and only if

$$\sum_{i,j=1}^{n} x_i\bar{x}_j \log \frac{1}{1-z_i\bar{z}_j} \geq \left| \sum_{i,j=1}^{n} x_i x_j \log \left[\frac{z_i z_j}{f(z_i)f(z_j)} \frac{f(z_i)-f(z_j)}{z_i - z_j} \right] \right| \tag{4.4.1}$$

for all choices of points $z_1, \ldots, z_n \in \mathbf{C}$ with $|z_i|<1$, all choices of points $x_1, \ldots, x_n \in \mathbf{C}$, and all $n=1,2,\ldots$. If $z_i=z_j$, the difference quotient on the right-hand side is to be interpreted as $f'(z_i)$. These formidable inequalities, known as the *Grunsky inequalities*, have the very simple algebraic form

$$x^*Ax \geq |x^TBx| \tag{4.4.2}$$

where $x=[x_i]\in\mathbf{C}^n$, $A=[a_{ij}]\in M_n$, $B=[b_{ij}]\in M_n$,

$$a_{ij} = \log\frac{1}{1-z_i\bar{z}_j}, \qquad b_{ij} = \log\left[\frac{z_i z_j}{f(z_i)f(z_j)} \frac{f(z_i)-f(z_j)}{z_i-z_j}\right]$$

Notice that A is Hermitian and B is complex symmetric.

Another example in which complex symmetric matrices arise naturally is in the general area of moment problems. Let $\{a_0, a_1, a_2, \ldots\}$ be a given sequence of complex numbers, let $n\geq 1$ be a given positive integer, and define $A_{2n}=[a_{ij}]\equiv[a_{i+j}]\in M_{2n}$. Notice that A_{2n} is a complex symmetric matrix of a form known as a *Hankel matrix*. We consider the complex quadratic form $x^TA_{2n}x$ for $x\in\mathbf{C}^{2n}$ and ask whether there is some fixed constant $c>0$ such that

$$|x^TA_{2n}x| \leq cx^*x \quad \text{for all} \quad x\in\mathbf{C}^{2n} \quad \text{for all} \quad n=1,2,\ldots$$

According to a theorem of Nehari, this condition is satisfied if and only if there is a Lebesgue measurable and almost everywhere bounded function $F(t): \mathbf{R}\to\mathbf{C}$ whose Fourier coefficients are the given numbers $a_0, a_1, a_2, \ldots$; the essential bound on $F(t)$ is exactly the constant c in the preceding inequalities.

Complex symmetric matrices do not seem to occur in applications nearly as often as complex Hermitian (or real symmetric) matrices, but the preceding examples show that they *do* occur. Although a complex symmetric matrix need not be diagonalizable (see Problem 15 at the end of this section), there is a factorization for complex symmetric matrices that is analogous to the spectral theorem (4.1.5) for Hermitian matrices, and it can be proved in a logically similar way. We first prove an analog of Schur's triangularization theorem (2.3.1) to show that a class of matrices that includes the symmetric matrices can always be factored as $A=U\Delta U^T$, where U is unitary and Δ is upper triangular. If an upper triangular matrix is symmetric, it must be diagonal.

4.4.3 Theorem. Let $A\in M_n$ be given. There exists a unitary $U\in M_n$ and an upper triangular $\Delta\in M_n$ such that $A=U\Delta U^T$ if and only if all the eigenvalues of $A\bar{A}$ are real and nonnegative. Under this condition, all the main diagonal entries of Δ may be chosen to be nonnegative.

Proof: If $A=U\Delta U^T$, then $A\bar{A}=U\Delta U^T\bar{U}\bar{\Delta}U^*=U\Delta\bar{\Delta}U^*$ since $\bar{U}$ is unitary and $U^T=\bar{U}^*$. The main diagonal entries of the upper triangular matrix $\Delta\bar{\Delta}$ are nonnegative real numbers whenever Δ is upper triangular, and $A\bar{A}$ is unitarily similar to $\Delta\bar{\Delta}$, so the necessity of the condition follows from the fact that the eigenvalues of an upper triangular matrix are exactly its main diagonal entries.

For the converse, assume that $A\bar{A}$ has only nonnegative eigenvalues, and let x be an eigenvector of $A\bar{A}$; that is, $A\bar{A}x=\lambda x$ with $\lambda\geq 0$ and $x\neq 0$. There are two possibilities:

(a) $A\bar{x}$ and x are dependent; or
(b) $A\bar{x}$ and x are independent.

In the former case (a) (which always happens when λ is a simple eigenvalue of $A\bar{A}$), there is some $\mu\in\mathbf{C}$ such that $A\bar{x}=\mu x$. But then $A\bar{A}x = A\bar{\mu}\bar{x}=\bar{\mu}A\bar{x}=\bar{\mu}\mu x=|\mu|^2x=\lambda x$, so $|\mu|^2=\lambda$. In the latter case (b) (which could happen if λ is a multiple eigenvalue of $A\bar{A}$), the vector $y=A\bar{x}+\mu x$ is nonzero for all $\mu\in\mathbf{C}$, and we choose μ to be any complex number such that $|\mu|^2=\mu\bar{\mu}=\lambda$. Then $A\bar{y}=A(\bar{A}x+\bar{\mu}\bar{x})=A\bar{A}x+\bar{\mu}A\bar{x}=\lambda x+\bar{\mu}A\bar{x}=\mu\bar{\mu}x+\bar{\mu}A\bar{x}=\bar{\mu}(A\bar{x}+\mu x)=\bar{\mu}y$. In either case (a) or (b), we have shown there is some nonzero vector $v\in\mathbf{C}^n$ and some $a\in\mathbf{C}$ with $|a|^2=\lambda$ such that $A\bar{v}=av$. Since this identity is unchanged if v is multiplied by a positive scalar, we may also assume that v is a unit vector. Also, for any $\theta\in\mathbf{R}$ we have $e^{-i\theta}A\bar{v}=A(\overline{e^{i\theta}v})=e^{-i\theta}av=(e^{-2i\theta}a)(e^{i\theta}v)$, and $e^{i\theta}v$ is a unit vector if v is. Since we can choose θ so that $e^{-2i\theta}a\geq 0$, we conclude that if $A\in M_n$, and if λ is a nonnegative eigenvalue of $A\bar{A}$, then there exists a unit vector v such that $A\bar{v}=\sigma v$, and $\sigma=+\sqrt{\lambda}\geq 0$.

Now extend this vector v to an orthonormal basis $\{v, v_2, \ldots, v_n\}$ of $\mathbf{C}^n$, and let V_1 be the unitary matrix that has these vectors as columns. The first column of the matrix $\bar{V}_1^TA\bar{V}_1$ has entries $v_i^*A\bar{v}=\sigma v_i^*v=\sigma\delta_{i1}$ because of orthonormality and the relation $A\bar{v}=\sigma v$. Thus, all but the first of the entries in the first column of $\bar{V}_1^TA\bar{V}_1$ must be zero (the first entry might also be zero). If we write this matrix in partitioned form as

$$\bar{V}_1^T A \bar{V}_1 = \begin{bmatrix} \sigma & w^T \\ 0 & A_2 \end{bmatrix}, \qquad w \in \mathbf{C}^{n-1}, \qquad A_2 \in M_{n-1}, \qquad \sigma \geq 0 \qquad (4.4.3a)$$

we see that

$$(\bar{V}_1^T A \bar{V}_1)(\overline{\bar{V}_1^T A \bar{V}_1}) = V_1^* A \bar{A} V_1 = \begin{bmatrix} \sigma^2 & \sigma \bar{w}^T + w^T \bar{A}_2 \\ 0 & A_2 \bar{A}_2 \end{bmatrix}$$

The eigenvalues of $A\bar{A}$ (all nonnegative by assumption) are therefore σ^2 together with the eigenvalues of $A_2\bar{A}_2$. We conclude that the matrix $A_2 \in M_{n-1}$ obtained by this process of reduction also has the property that all the eigenvalues of $A_2\bar{A}_2$ are nonnegative.

The process of reduction can now be repeated with A_2 and its successors at most $n-1$ times [just as in the proof of the Schur triangularization theorem (2.3.1)] to obtain

$$\bar{V}_{n-1}^T \cdots \bar{V}_2^T \bar{V}_1^T A \bar{V}_1 \bar{V}_2 \cdots \bar{V}_{n-1} = \begin{bmatrix} \sigma_1 & & * \\ & \ddots & \\ 0 & & \sigma_n \end{bmatrix} = \Delta$$

where Δ is upper triangular with nonnegative main diagonal entries σ_i. If we set $U = V_1 V_2 \cdots V_{n-1}$, we have $A = U\Delta U^T$, as desired. □

Exercise. Explicitly carry out the calculations in the proof of Theorem (4.4.3) for the matrix $A = \left[\begin{smallmatrix} 1 & i \\ -i & 1 \end{smallmatrix}\right]$ and show that $A = U\Delta U^T$ with

$$\Delta = \begin{bmatrix} 0 & 2i \\ 0 & 0 \end{bmatrix}, \qquad U = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}$$

If $n \geq 2$, not every matrix $A \in M_n$ has the property that $A\bar{A}$ has all nonnegative eigenvalues; $A = \left[\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right]$ is a simple example. Thus, Theorem (4.4.3) is only a partial analog of Schur's triangularization theorem (2.3.1). Every $A \in M_n$ can be triangularized by a transformation of the form $A \to UAU^*$ with a unitary $U \in M_n$, but only those matrices $A \in M_n$ such that $A\bar{A}$ has nonnegative eigenvalues can be triangularized by a transformation of the form $A \to UAU^T$ with a unitary $U \in M_n$.

Every symmetric matrix $A \in M_n$ has the property that all the eigenvalues of $A\bar{A} = AA^*$ are nonnegative, however. The special form that Theorem (4.4.3) takes in this case is commonly attributed to Schur (1945), but earlier proofs were offered by Hua (1944), Siegel (1943), and Jacobsen (1939); historical priority must apparently be given to Takagi (1925).

4.4.4 **Corollary** (Takagi's factorization). If $A \in M_n$ is symmetric ($A = A^T$), then there exists a unitary $U \in M_n$ and a real nonnegative diagonal

matrix $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n)$ such that $A = U\Sigma U^T$. The columns of U are an orthonormal set of eigenvectors for $A\bar{A}$, and the corresponding diagonal entries of Σ are the nonnegative square roots of the corresponding eigenvalues of $A\bar{A}$.

Proof: If $A = A^T$, then $\bar{A} = A^*$ and $A\bar{A} = AA^*$. If $x \neq 0$ is any eigenvector of the Hermitian matrix AA^* and $AA^*x = \lambda x$, then $x^*\lambda x = \lambda(x^*x) = x^*AA^*x = (A^*x)^*(A^*x)$. Since $y^*y \geq 0$ for all $y \in \mathbf{C}^n$ with $y^*y = 0$ if and only if $y = 0$, we see that $\lambda = (A^*x)^*(A^*x)/x^*x \geq 0$. Thus, all the eigenvalues of $A\bar{A}$ are nonnegative whenever A is symmetric. The theorem guarantees that there is a unitary $U \in M_n$ and an upper triangular $\Delta \in \mathbf{M}_n$ with

$$\Delta = \begin{bmatrix} \sigma_1 & & * \\ & \ddots & \\ 0 & & \sigma_n \end{bmatrix}, \qquad \text{all } \sigma_i \geq 0$$

such that $A = U\Delta U^T$. But then $U\Delta U^T = A = A^T = U\Delta^T U^T$, so $\Delta = \Delta^T$, which can happen only if $\Delta \equiv \Sigma$ is a diagonal matrix, which is nonnegative by construction. Finally, $A\bar{A} = U\Sigma U^T \bar{U}\Sigma U^* = U\Sigma^2 U^*$ is a unitary diagonalization of the Hermitian matrix $A\bar{A}$, so the columns of U are eigenvectors of $A\bar{A}$. □

Any matrix of the form $U\Lambda U^T$ with Λ diagonal (not necessarily nonnegative) is evidently symmetric, so in order that a given matrix $A \in M_n$ can be factored as $A = U\Lambda U^T = U\Lambda\bar{U}^* = U\Lambda\bar{U}^{-1}$, with U unitary and Λ diagonal, it is necessary and sufficient that A be symmetric. Conditions under which A can be factored as $A = S\Lambda\bar{S}^{-1}$ with Λ diagonal and S nonsingular (but not necessarily unitary) are given in Theorem (4.6.11).

Every complex matrix $A \in M_n$ can be written in the form $A = V\Sigma W^*$, where $V, W \in M_n$ are unitary and Σ is a diagonal matrix with nonnegative main diagonal entries. This is the *singular value decomposition,* and it is discussed in Section (7.3). The diagonal entries of Σ are the *singular values* of A. The Takagi factorization $A = U\Sigma U^T$ for a (possibly complex) symmetric matrix is a special singular value decomposition for symmetric matrices in which $V = \bar{W}$.

The construction used in the proof of Theorem (4.4.3) can be used to compute a Takagi factorization of a complex symmetric matrix. The matrix Δ produced will automatically be diagonal because of the symmetry of A. See Problem 9 at the end of this section.

Exercise. Explicitly carry out the calculations in the proof of Theorem (4.4.3) for the matrix $A = \left[\begin{smallmatrix} 1 & i \\ i & 1 \end{smallmatrix}\right]$ and show that $A = U\Delta U^T$ with

$$\Delta=\begin{bmatrix}\sqrt{2} & 0\\ 0 & \sqrt{2}\end{bmatrix} \quad\text{and}\quad U=\frac{1}{\sqrt{4+2\sqrt{2}}}\begin{bmatrix}1+\sqrt{2} & i\\ i & 1+\sqrt{2}\end{bmatrix}$$

Notice that Δ is automatically diagonal.

Because the columns of the unitary factor U in the Takagi factorization $A=U\Sigma U^T$ are eigenvectors of the Hermitian matrix $A\bar{A}$, there may be a temptation to assume that if $A\bar{A}=U\Sigma^2U^*$ is a unitary diagonalization, then $A=U\Sigma U^T$. This need not be the case, as can be seen from consideration of the example $A=\left[\begin{smallmatrix}0 & 1\\ 1 & 0\end{smallmatrix}\right]$. Since $A\bar{A}=I$, we have $A\bar{A}=QI^2Q^T$ for *any* real orthogonal 2-by-2 matrix Q, but $QIQ^T=I\neq A$. The problem is that $A\bar{A}$ has an eigenvalue with multiplicity greater than 1, so an arbitrary eigenvector x of $A\bar{A}$ might not have the property that $A\bar{x}=ax$; this eigenvector might not give the desired reduction of A. If we consider the basis vector e_1, then $A\bar{A}e_1=Ie_1=1e_1$, but $A\bar{e}_1=Ae_1=e_2$; we have case (b) in the proof of Theorem (4.4.3). According to the proof, we may take $w=A\bar{e}_1+1e_1=e_2+e_1$ to obtain a vector $v=v_1=(e_1+e_2)/\sqrt{2}$, which will reduce A. Since $v_2=(e_1-e_2)/\sqrt{2}$ is orthogonal to v_1, we may take

$$V=\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}$$

and obtain $V^TAV=\left[\begin{smallmatrix}1 & 0\\ 0 & -1\end{smallmatrix}\right]=\left[\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right]\left[\begin{smallmatrix}1 & 0\\ 0 & i\end{smallmatrix}\right]^2=\Sigma D^2$. Thus, if we set

$$U=VD=\frac{1}{\sqrt{2}}\begin{bmatrix}1 & i\\ 1 & -i\end{bmatrix}$$

$A=UIU^T$ is an appropriate factorization of A. Notice that a Takagi factorization (4.4.4) of a real symmetric matrix might not have real factors.

The difficulty in the example just discussed, and in the general case, arises from multiple eigenvalues of $A\bar{A}$. If all the eigenvalues of $A\bar{A}$ are distinct, and if one uses the construction in the proof of (4.4.3) to compute a Takagi factorization of the complex symmetric matrix A, one always has case (a). (See Problem 9.) In this case, each eigenvector x of $A\bar{A}$ has the property that $A\bar{x}=ax$ for some $a\in\mathbf{C}$ such that $a=\sigma e^{2i\theta}$, $\theta\in\mathbf{R}$, and $A\bar{A}x=\sigma^2x$. Thus, if $A\bar{A}=V\Sigma^2V^*$ is a unitary diagonalization of the Hermitian matrix $A\bar{A}$, we must have $A\bar{V}=V\Sigma D^2$, where $D^2\equiv\operatorname{diag}(e^{2i\theta_1},\ldots,e^{2i\theta_n})$; this identity can be used to compute the diagonal entries of D^2 corresponding to nonzero diagonal entries of Σ once V and Σ (the nonnegative square root of Σ^2) are known. The entries of D^2 corresponding to zero entries of Σ are arbitrary and may be taken to be $+1$. Finally, we have $A=A\bar{V}V^T=V\Sigma D^2V^T=(VD)\Sigma(VD)^T=U\Sigma U^T$ if we set $U\equiv VD$ and $D=\operatorname{diag}(e^{i\theta_1},\ldots,e^{i\theta_n})$. We record these observations formally as the following corollary.

4.4.5 Corollary. If $A\in M_n$ is symmetric, if the eigenvalues of $A\bar{A}$ are distinct, and if $A\bar{A}=V\Sigma^2V^*$ is a unitary diagonalization of $A\bar{A}$ with $\Sigma=\operatorname{diag}(\sigma_1,\ldots,\sigma_n)$ and all $\sigma_i\geq 0$, then there exists a diagonal matrix $D=\operatorname{diag}(e^{i\theta_1},\ldots,e^{i\theta_n})$ with all $\theta_j\in\mathbf{R}$ such that $A=U\Sigma U^T$ with $U=VD$. The diagonal entries of the factor D corresponding to nonzero diagonal entries of Σ are determined by the relation $A\bar{V}=V\Sigma D^2$; the diagonal entries of D corresponding to zero diagonal entries of Σ may be taken to be $+1$.

If $A\in M_n$ is symmetric, and if we use (4.4.4) to write $A=U\Sigma U^T$, then we can also write this as $A=(U\Sigma^{1/2})(U\Sigma^{1/2})^T$, where $\Sigma^{1/2}=\operatorname{diag}(+\sqrt{\sigma_1},+\sqrt{\sigma_2},\ldots,+\sqrt{\sigma_n})$. This observation constitutes a proof of the following corollary.

4.4.6 Corollary. Let $A\in M_n$. Then A is symmetric if and only if there exists a matrix $S\in M_n$ such that $A=SS^T$. One may choose $S=UD$ where U is unitary, $D=\operatorname{diag}(\sqrt{\sigma_1},\sqrt{\sigma_2},\ldots,\sqrt{\sigma_n})$, and $\{\sigma_i\}$ are the singular values of A, in which case rank $S=$ rank A.

Although a real symmetric matrix is normal, a nonreal complex symmetric matrix need not be normal. If $A=B+iC\in M_n$ with B and C real, then A is symmetric if and only if B and C are both real symmetric matrices. If A is both symmetric and normal, then

$$AA^*=(B^2+C^2)+i(CB-BC)=(B^2+C^2)+i(BC-CB)=A^*A$$

from which it follows that B and C commute. In this event, B and C are simultaneously diagonalizable by a real orthogonal matrix Q. If $B=QD_1Q^T$ and $C=QD_2Q^T$ with D_1 and D_2 real diagonal matrices, then $A=B+iC=QD_1Q^T+iQD_2Q^T=Q(D_1+iD_2)Q^T=Q\Lambda Q^T$ with $\Lambda=D_1+iD_2$. Conversely, if a matrix $A\in M_n$ can be written as $A=Q\Lambda Q^T$ with Q a real orthogonal matrix and Λ a diagonal matrix, then $A=A^T$ and $AA^*=Q\Lambda Q^TQ\bar{\Lambda}Q^T=Q|\Lambda|^2Q^T=Q\bar{\Lambda}Q^TQ\Lambda Q^T=A^*A$, so A is both symmetric and normal. This proves the following theorem.

4.4.7 Theorem. Let $A\in M_n$. Then A is both symmetric and normal if and only if there is a real orthogonal matrix $Q\in M_n(\mathbf{R})$ and a diagonal matrix $\Lambda\in M_n$ such that $A=Q\Lambda Q^T$.

A useful example of a simple complex matrix that is both symmetric and normal is

$$S=\frac{1}{\sqrt{2}}(I+iB) \tag{4.4.8}$$

where B is the "backward identity" matrix

$$B = \begin{bmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{bmatrix}$$

which played a role in (3.2.3) in showing that every matrix is similar to its transpose.

Since $B^2 = I$,

$$S\bar{S} = \tfrac{1}{2}(I + iB)(I - iB) = \tfrac{1}{2}(I - iB + iB + B^2) = I$$

and we see that S is both symmetric and unitary.

Now consider a typical Jordan block $J_k(0)$ with zero main diagonal and $k \geq 2$, which we write in the form

$$N = \begin{bmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{bmatrix} \in M_k$$

It is a simple computation to show that

$$BNB = \begin{bmatrix} 0 & & & 0 \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ 0 & & 1 & 0 \end{bmatrix}$$

$$BN = \begin{bmatrix} 0 & & & 0 \\ & & \iddots & 1 \\ & \iddots & \iddots & \\ 0 & 1 & & 0 \end{bmatrix}$$

$$NB = \begin{bmatrix} 0 & & 1 & 0 \\ & \iddots & \iddots & \\ 1 & \iddots & & \\ 0 & & & 0 \end{bmatrix}$$

Thus, N is unitarily similar to the matrix

$$SNS^{-1} = SN\bar{S} = \frac{1}{2}(I + iB)N(I - iB)$$

$$= \frac{1}{2}(N + BNB) + \frac{i}{2}(BN - NB)$$

$$= \frac{1}{2}\begin{bmatrix} 0 & 1 & & 0 \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & 0 \end{bmatrix} + \frac{i}{2}\begin{bmatrix} 0 & & -1 & 0 \\ & \iddots & \iddots & 1 \\ -1 & \iddots & \iddots & \\ 0 & 1 & & 0 \end{bmatrix} \tag{4.4.8a}$$

which is evidently symmetric. Any Jordan block $J_k(\lambda)$ with $k \geq 2$ is of the form $\lambda I + N$, and $SJ_k(\lambda)S^{-1} = S(\lambda I + N)S^{-1} = \lambda I + SNS^{-1}$ is symmetric since SNS^{-1} is symmetric.

Every matrix $A \in M_n$ is similar to a Jordan canonical form J of the form (3.1.14) with $\epsilon = 2$, and $J = J_{n_1}(\lambda_1, 2) \oplus \cdots \oplus J_{n_k}(\lambda_k, 2)$ is a direct sum of modified Jordan blocks $J_{n_i}(\lambda_i, 2)$. This observation (equivalent to replacing N by $2N$ in the preceding argument) permits us to drop the coefficient factors of $\frac{1}{2}$ in (4.4.8a). If we let $S_{n_i} \equiv (1/\sqrt{2})(I + iB) \in M_{n_i}$ be the n_i-by-n_i matrix of the form (4.4.8) if $n_i \geq 2$ and $S_1 \equiv [1]$, and if we set $T = S_{n_1} \oplus \cdots \oplus S_{n_k}$, then the preceding argument shows that

$$TJT^{-1} = TJ\bar{T} = (S_{n_1}J_{n_1}(\lambda_1, 2)\bar{S}_{n_1}) \oplus \cdots \oplus (S_{n_k}J_{n_k}(\lambda_k, 2)\bar{S}_{n_k})$$

is a direct sum of symmetric matrices and is therefore symmetric. The matrix T is unitary since each S_{n_i} is unitary, so we have shown that every matrix in Jordan canonical form is unitarily equivalent to a symmetric matrix. Since every matrix is similar to a Jordan matrix, we have proved the following theorem.

4.4.9 **Theorem.** Every matrix $A \in M_n$ is similar to a symmetric matrix.

In fact, we have shown that every matrix $A \in M_n$ is similar to a *symmetric Jordan canonical form* $S_{n_1}(\lambda_1) \oplus \cdots \oplus S_{n_k}(\lambda_k)$, where

$$S_k(\lambda) = SJ_k(\lambda, 2)\bar{S} = \lambda I + SN\bar{S}$$

$$= \lambda I + \begin{bmatrix} 0 & 1 & & 0 \\ 1 & 0 & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & 0 \end{bmatrix} + i\begin{bmatrix} 0 & & -1 & 0 \\ & \iddots & \iddots & 1 \\ -1 & \iddots & \iddots & \\ 0 & 1 & & 0 \end{bmatrix} \in M_k$$

and S is given by (4.4.8). Note that

$$S_1(\lambda) = [\lambda] \quad \text{and} \quad S_2(\lambda) = \begin{bmatrix} \lambda - i & 1 \\ 1 & \lambda + i \end{bmatrix}$$

Since this form was derived from the Jordan canonical form, its uniqueness is the same as that of the Jordan canonical form.

One of the consequences of this result is that there is nothing special about the spectrum, Jordan blocks, minimal polynomial, characteristic polynomial, or invariant factors of a symmetric complex matrix. Any of these quantities can occur for a symmetric matrix of a given size that can occur for a general complex matrix of the same size. Every similarity class in M_n contains a symmetric matrix, every linear transformation on $\mathbf{C}^n$ has a symmetric basis representation, and symmetry of a matrix is just an artifact of the particular basis chosen to represent the underlying linear transformation. Another consequence is that every matrix is "diagonalizable" in a certain sense.

4.4.10 **Corollary.** Let $A \in M_n$ be given. There is a nonsingular matrix S and a unitary matrix U such that $(US)A(\bar{U}S)^{-1}$ is a diagonal matrix with nonnegative diagonal entries.

Proof: Use (4.4.9) to find a nonsingular $S \in M_n$ such that SAS^{-1} is symmetric and then use (4.4.4) to find a unitary $U \in M_n$ such that $U(SAS^{-1})U^T$ is diagonal and nonnegative. $\square$

Theorem (4.4.9) also implies that every complex matrix is similar to its transpose and can be written as a product of two complex symmetric matrices. Both of these results are true for matrices over any field, but Theorem (4.4.9) is not true for general fields.

4.4.11 **Corollary.** Let $A \in M_n$ be given. There are matrices $B, C \in M_n$ such that $B = B^T$, $C = C^T$, and $A = BC$. Either B or C may be chosen to be nonsingular.

Proof: Use the theorem to write $A = SES^{-1}$, where $E = E^T$ and S is nonsingular. Then $A = (SES^T)(S^T)^{-1}S^{-1} = (SES^T)(SS^T)^{-1} = BC$, where $B = SES^T$ and $C = SS^T$ are both symmetric. Since $A = (SS^T)(S^{-1})^T ES^{-1}$ as well, either factor B or C may be chosen to be nonsingular. $\square$

The Gram–Schmidt process (0.6.4) has many applications in the study of normal matrices. There is an analogous process that is useful in the study of complex symmetric matrices.

4.4.12 **Lemma.** Let $x_1, \ldots, x_k \in \mathbf{C}^n$ be given vectors with $k \le n$. There exist vectors $y_1, \ldots, y_k$ such that $\text{Span}\{x_1, \ldots, x_k\} = \text{Span}\{y_1, \ldots, y_k\}$, $y_i^T y_j = 0$ for all $i, j = 1, 2, \ldots, k$ with $i \ne j$, $y_i^T y_i = 1$ for $i = 1, 2, \ldots, r$, and $y_i^T y_i = 0$ for $i = r+1, \ldots, k$, where $r = \text{rank}\, X^T X$ and $X = [x_1 \cdots x_k] \in M_{n,k}$ is the matrix whose columns are the given vectors $\{x_i\}$.

Proof: Because the matrix $X^T X$ is symmetric, Takagi's factorization theorem (4.4.4) permits us to write $X^T X = U \Sigma U^T$, where $U \in M_k$ is unitary and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_k)$ with $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > \sigma_{r+1} = 0 = \cdots = \sigma_k$, and $\text{rank}\, X^T X = r$. If we set $D = \text{diag}(\sqrt{\sigma_1}, \ldots, \sqrt{\sigma_r}, 1, \ldots, 1) \in M_k$, and write $I_r = \text{diag}(1, \ldots, 1, 0, \ldots, 0) \in M_k$ with r 1's and $k-r$ 0's, then $X^T X = (UD) I_r (UD)^T = S^T I_r S$, where $S = DU^T$ is nonsingular. Thus, $(XS^{-1})^T (XS^{-1}) = I_r$, so if we set $XS^{-1} \equiv Y = [y_1 \cdots y_k] \in M_{n,k}$, the column vectors $y_1, \ldots, y_k$ have the asserted properties since $Y^T Y = I_r$. $\square$

The preceding lemma states a principle that is formally similar to the Gram–Schmidt process, which involves X^*X instead of $X^T X$. In the Gram–Schmidt process, however, each y_j can be formed as a linear combination of $x_1, \ldots, x_j$ for each $j = 1, 2, \ldots, k$ and that may not be possible here. Another difference is that the number of vectors y_i for which $y_i^* y_i = 1$ in the Gram–Schmidt process is equal to $\text{rank}\, X$ (the maximum number of independent vectors x_i), which is always equal to $\text{rank}\, X^*X$. In this case, however, the number of vectors y_i for which $y_i^T y_i = 1$ is equal to $\text{rank}\, X^T X$, which can be less than $\text{rank}\, X$.

Example. Consider $k = 1$ and $x_1 = X = \left[\begin{smallmatrix} 1 \\ i \end{smallmatrix}\right]$. Then $X^T X = 0$, so $0 = \text{rank}\, X^T X$, which is strictly less than $\text{rank}\, X = 1$. The only possibility for y_1 is a scalar multiple of x_1, so it is not possible to choose y_1 so that $\text{Span}\{x_1\} = \text{Span}\{y_1\}$ and $y_1^T y_1 = 1$.

Example. Consider $k = 2$ and $X = [x_1\ x_2] = \left[\begin{smallmatrix} 1 & i \\ i & 1 \end{smallmatrix}\right]$. Then $\text{rank}\, X^T X = 2$ and there exist vectors y_1, y_2 such that $\text{Span}\{y_1, y_2\} = \text{Span}\{x_1, x_2\}$ and $y_1^T y_1 = 1 = y_2^T y_2$. Since $x_1^T x_1 = 0$, it is not possible to choose y_1 to be a scalar multiple of x_1.

The immediate application we have in mind is to the special situation of diagonalizable complex symmetric matrices. If $A = A^T \in M_n$ and if $A = S \Lambda S^{-1}$ for a diagonal $\Lambda \in M_n$ and a nonsingular $S \in M_n$, then it is not evident from this usual diagonalization representation that A is symmetric. If S is a complex orthogonal matrix, however, then $S^{-1} = S^T$ and $A = S \Lambda S^{-1} = S \Lambda S^T$ is evidently symmetric. The next theorem says that it is always possible to choose S to be complex orthogonal.

4.4.13 **Theorem.** Let $A \in M_n$ be symmetric. Then A is diagonalizable if and only if it is complex orthogonally diagonalizable, that is, $A = S \Lambda S^{-1}$

for a diagonal $\Lambda \in M_n$ and a nonsingular $S \in M_n$, if and only if $A = Q\Lambda Q^T$, where $Q \in M_n$ satisfies $Q^TQ = I$.

Proof: Suppose $A = A^T$ and let $x, y \in \mathbf{C}^n$ be eigenvectors of A with $Ax = \lambda x$ and $Ay = \mu y$. If $\lambda \neq \mu$, then $y^TAx = y^T\lambda x = \lambda y^Tx$ and $y^TAx = (Ay)^Tx = (\mu y)^Tx = \mu y^Tx$. Thus $\lambda y^Tx = \mu y^Tx$ and $y^Tx = 0$ since $\lambda \neq \mu$. This is just an application of the principle of biorthogonality (1.4.7) to symmetric matrices. If A is diagonalizable and $A = S\Lambda S^{-1}$, there is no loss of generality to assume that the like eigenvalues of A are grouped together in $\Lambda = \Lambda_1 \oplus \cdots \oplus \Lambda_d$ with $\Lambda_i = \lambda_i I \in M_{n_i}$, $n_1 + \cdots + n_d = n$ and $\lambda_i \neq \lambda_j$ if $i \neq j$. Partition the columns of $S = [s_1 \cdots s_n] = [S_1\ S_2 \cdots S_d]$ conformally with $\Lambda = \Lambda_1 \oplus \cdots \oplus \Lambda_d$, so that $S_i \in M_{n,n_i}$ for $i = 1, 2, \ldots, d$. Because of the biorthogonality property, $S_i^TS_j = 0 \in M_{n_i,n_j}$ if $i \neq j$, and $S_i^TS_i$ is nonsingular for all $i = 1, 2, \ldots, d$ because S^TS is nonsingular and block diagonal. Since each matrix $S_i^TS_i$ has full rank, Lemma (4.4.12) says that the columns of each S_i can be replaced by new columns, which are a nonsingular linear combination of the old ones and are mutually complex orthogonal; that is, there exists a nonsingular $R_i \in M_{n_i}$ such that $Q_i \equiv S_iR_i$ satisfies $Q_i^TQ_i = R_i^TS_i^TS_iR_i = I \in M_{n_i}$. Since $Q_i^TQ_j = R_i^TS_i^TS_jR_j = 0$ for all $i \neq j$ and $AQ_i = AS_iR_i = \lambda_iS_iR_i = \lambda_iQ_i$ for $i = 1, 2, \ldots, d$, the matrix $Q = [Q_1 \cdots Q_d] \in M_n$ is complex orthogonal and $A = Q\Lambda Q^T$. $\square$

The preceding result provides an interesting setting for Theorem (4.4.7): A symmetric matrix A is diagonalizable if and only if $A = Q\Lambda Q^T$ with Q complex orthogonal, and it is normal if and only if Q can be chosen to be real.

The result in Theorem (4.4.13) can be generalized somewhat. If $A, B \in M_n$ are symmetric matrices, then A and B are similar if and only if they are similar via a complex orthogonal similarity. In fact, this is true under the weaker hypothesis that there is one polynomial $p(t)$ such that $A^T = p(A)$ and $B^T = p(B)$. See [HJ].

Problems

1. Suppose $A \in M_n$ is symmetric and $A = B + iC$ with $B, C \in M_n$ both real. Show that A is normal if and only if B and C commute. Show that A is normal if and only if $A\bar{A}$ is real. Show that A is normal if and only if A and $\bar{A}$ commute. Give an example of a symmetric matrix that is not normal.

2. Provide the details for the following outline for a different proof of Corollary (4.4.4). The notation and hypotheses are as in (4.4.4). If A is singular, let $\{u_1, \ldots, u_k\}$ be an orthonormal basis of the null space of A and let $U = [u_1 \ldots u_k\ u_{k+1} \ldots u_n] \in M_n$ be unitary. Then

$$U^TAU = \begin{bmatrix} 0 & 0 \\ 0 & A' \end{bmatrix}, \quad A' \in M_{n-k}$$

with A' nonsingular and symmetric. Thus, without loss of generality, we may assume that A is nonsingular. Let $A = B + iC$ with B, C real and let $z = x + iy \in \mathbf{C}^n$ with $x, y \in \mathbf{R}^n$. Let $F = \left[\begin{smallmatrix} B & C \\ C & -B \end{smallmatrix}\right]$, $\tilde{z} = \left[\begin{smallmatrix} x \\ -y \end{smallmatrix}\right] \in \mathbf{R}^{2n}$. (a) B, C, and F are real symmetric matrices. Discuss the relationship between $Az = (B + iC)(x + iy)$ and $F\tilde{z}$. (b) F is nonsingular. *Hint:* If $F\tilde{z} = 0$, what is Az? (c) If $F\left[\begin{smallmatrix} x \\ -y \end{smallmatrix}\right] = \lambda\left[\begin{smallmatrix} x \\ -y \end{smallmatrix}\right]$, then $F\left[\begin{smallmatrix} y \\ x \end{smallmatrix}\right] = -\lambda\left[\begin{smallmatrix} y \\ x \end{smallmatrix}\right]$. The nonzero eigenvalues of F can be paired with one positive and one negative. (d) Let the orthonormal eigenvectors of F corresponding to the positive eigenvalues $\lambda_1, \ldots, \lambda_n$ be denoted $\tilde{z}_i = \left[\begin{smallmatrix} x_i \\ -y_i \end{smallmatrix}\right] \in \mathbf{R}^{2n}$, $i = 1, 2, \ldots, n$, let $X \equiv [x_1 \cdots x_n]$, $Y \equiv [y_1 \cdots y_n] \in M_n$, and let $\Sigma = \text{diag}(\lambda_1, \ldots, \lambda_n) \in M_n$. The spectral theorem for real symmetric matrices says that $F = V\Lambda V^T$, where

$$V = \begin{bmatrix} X & Y \\ -Y & X \end{bmatrix} \quad \text{and} \quad \Lambda = \begin{bmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{bmatrix}$$

and V is a real orthogonal matrix (why?). Let $U \equiv X - iY$. Show that U is unitary and that $U\Sigma U^T = A$.

3. What does (4.4.4) say when A is a real symmetric matrix? How is it related to the usual spectral decomposition of a real symmetric matrix? *Hint:* If $A = Q\Lambda Q^T$ with Λ a real diagonal matrix and Q a real orthogonal matrix, write $\Lambda = \Sigma D^2$ and let $U = QD$. When can all the factors in the Takagi factorization $A = U\Sigma U^T$ be taken to be real?

4. If $A = U\Sigma U^T \in M_n$ with U and Σ as in (4.4.4), show by direct computation that σ_i^2 are the eigenvalues of $\bar{A}A$ and $A\bar{A}$, and that $\bar{A}A$ and $A\bar{A}$ are Hermitian. Show that the columns u_i of U and the numbers σ_i satisfy the equations $A\bar{u}_i = \sigma_iu_i$, $i = 1, 2, \ldots, n$. Perhaps for this reason, the σ_i are sometimes called *generalized eigenvalues,* but the term *singular values* seems to be more common.

5. Let $A \in M_n$ be symmetric, let Σ and U be as in (4.4.4), and arrange the singular values of A in nonincreasing order $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$. (a) Modify the proof of the Rayleigh–Ritz theorem (4.2.2) to show that $\sigma_{\max} = \sigma_1 = \max\{|x^TAx|/x^*x : 0 \neq x \in \mathbf{C}^n\}$, that is, there is a complex symmetric analog of the upper bound in (4.2.2). Consider the first column of U to show that the extremum is obtained for a unit vector x such that $A\bar{x} = \sigma_1x$. (b) Consider $A = I \in M_2$ and $x = \left[\begin{smallmatrix} 1 \\ i \end{smallmatrix}\right]$ to show that $\sigma_{\min} = \sigma_n \neq \min\{|x^TAx|/x^*x : 0 \neq x \in \mathbf{C}^n\}$ in this case, so a complex symmetric analog

of the lower bound in (4.2.2) is false. (c) Consider $A = I \in M_2$, $w = \begin{bmatrix} 1 \\ i \end{bmatrix}$ and show that $\max\{|x^TAx|/x^*x : 0 \neq x \in \mathbf{C}^n,\ x \perp w\} = 0$. Conclude that a complex symmetric matrix/singular value analog of the Courant–Fischer min-max formula (4.2.12) is false for $k > 1$. However, see (7.3.10). (d) What about a symmetric analog of the max-min formula (4.2.13)? (e) Let $\tilde{A} \equiv \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}$ $(\tilde{\sigma}_1 = \tilde{\sigma}_2 = \sqrt{2})$ and form $A = [1]$ $(\sigma_1 = 1)$ by deleting the last row and column from $\tilde{A}$. Note that the interlacing inequality $\tilde{\sigma}_1 \geq \sigma_1 \geq \tilde{\sigma}_2$ analogous to (4.3.9) is not correct. (f) Nevertheless, there are inequalities for the singular values of bordered symmetric matrices. Let $\tilde{A} \in M_{n+1}$ be symmetric and have singular values $\tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_{n+1}$ and form $A \in M_n$ (with singular values $\sigma_1 \geq \cdots \geq \sigma_n$) by deleting a row and corresponding column from $\tilde{A}$. Use Theorem (7.3.9) to show that $\tilde{\sigma}_k \geq \sigma_k \geq \tilde{\sigma}_{k+2}$, $k = 1, \ldots, n$ $(\tilde{\sigma}_{n+2} \equiv 0)$. Verify these inequalities for the example in (e) and compare them with the interlacing inequalities (4.3.9) for eigenvalues of bordered Hermitian matrices.

6. If $A \in M_n$ is symmetric and if $A = U\Sigma U^T$ with U unitary and $\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ with all $\sigma_i \geq 0$, show that the rank of A is equal to the number of nonzero σ_i terms. *Hint:* If $B, C \in M_n$ are nonsingular, then rank A = rank BAC.

7. Let $A = B + iC \in M_n$ with B, C real, and let $F = \begin{bmatrix} B & C \\ C & -B \end{bmatrix} \in M_{2n}$. (a) Show that $\bar{A}A = B^2 + C^2 + i(BC - CB)$ and

$$F^2 = \begin{bmatrix} B^2 + C^2 & BC - CB \\ -(BC - CB) & B^2 + C^2 \end{bmatrix}$$

(b) Show that $S \equiv (1/\sqrt{2})\begin{bmatrix} I & -iI \\ -iI & I \end{bmatrix} \in M_{2n}$ is unitary. (c) Show that $SF^2S^* = \begin{bmatrix} \bar{A}A & 0 \\ 0 & A\bar{A} \end{bmatrix}$. (d) Conclude that the squares of the eigenvalues of F are the same as the eigenvalues of $\bar{A}A$, together with their complex conjugates. (e) If A is a complex symmetric matrix, show that F is a real symmetric matrix with real eigenvalues, that F^2 has only nonnegative eigenvalues, and that the set of squares of the eigenvalues of F is the same as the set of eigenvalues of the Hermitian matrix $\bar{A}A$.

8. Let $A \in M_n$ be a complex symmetric matrix. Consider the quadratic form $q_A(x, x) = x^TAx$ and the bilinear form $b_A(x, y) = x^TAy$ generated by A. Use Corollary (4.4.4) to show that

$$\sup_{x^*x=1} |q_A(x, x)| = \sup_{\substack{x^*x=1 \\ y^*y=1}} |b_A(x, y)| = \sigma_{\max}(A)$$

where $\sigma^2_{\max}(A)$ is the largest eigenvalue of $\bar{A}A$.

9. Using the notation of the proof of (4.4.3), show the following: (i) If λ is a simple eigenvalue of $A\bar{A}$ and $x \neq 0$ is such that $A\bar{A}x = \lambda x$, then possibility (a) is always the case. *Hint:* Let $\sigma = +\sqrt{\lambda}$ and set $w = A\bar{x} - \sigma x$. Show that $A\bar{w} = -\sigma w$ and $A\bar{A}w = \lambda w$, so w is a scalar multiple of x. (ii) If $A = A^T$, then $\bar{V}_1^TA\bar{V}_1 = [\sigma] \oplus A_2$; that is, the row vector w^T in (4.4.3a) is zero. Use this to show that the construction automatically produces a matrix $\bar{V}_{n-1}^T \cdots \bar{V}_1^TA\bar{V}_1\bar{V}_2 \cdots \bar{V}_{n-1} = U^*A\bar{U} = \Delta$ that is diagonal.

10. Let $A \in M_n$ and suppose there is a nonsingular $S \in M_n$ such that $A = S\Lambda\bar{S}^{-1}$, where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Show that $A\bar{A}$ is diagonalizable and has only nonnegative eigenvalues, and that rank A = rank $A\bar{A}$. What does this have to do with (4.4.4)? Show that neither $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ nor $\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ can be written in this form.

11. If $S \in M_n$ is given, show that rank $S^TS \leq$ rank S in general and that rank $S^TS <$ rank S is possible. What happens if S is real? *Hint:* Consider $S = \begin{bmatrix} 1 & 0 \\ i & 0 \end{bmatrix}$.

12. If $A \in M_n$ is a complex symmetric matrix and if $x, y \in \mathbf{C}^n$ are eigenvectors of A corresponding to distinct eigenvalues of A, show that $x^Ty = 0$. Does this mean that x and y are orthogonal? *Hint:* Consider $x^T(Ay) = (Ax)^Ty$.

13. If $A \in M_n$ is symmetric and has n distinct eigenvalues, show directly that there exists a nonsingular matrix $S \in M_n$ and a diagonal matrix D such that $A = SDS^T$. *Hint:* A is diagonalizable, so $A = S\Lambda S^{-1}$ and $AS = S\Lambda$. By Problem 12, $S^TS = D$ is diagonal. Thus $S^TAS = S^TS\Lambda = D\Lambda$ and $A = (S^{-1})^T\ (D\Lambda)S^{-1}$. What adjustments need to be made to show that $A = Q\Lambda Q^T$ with a complex orthogonal Q?

14. If $A \in M_n$ is symmetric and nonsingular, show that A^{-1} is symmetric.

15. A real symmetric matrix is Hermitian and therefore is diagonalizable. Show that a complex symmetric matrix need not be diagonalizable. *Hint:* Consider $A = \begin{bmatrix} 1 & i \\ i & -1 \end{bmatrix}$ and compute A^2.

16. Let $A \in M_n$. Show that A is both symmetric and unitary if and only if A can be written as $A = Q\Lambda Q^T$, where $Q \in M_n(\mathbf{R})$ is a real orthogonal matrix and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) = \operatorname{diag}(e^{i\theta_1}, \ldots, e^{i\theta_n})$ with $|\lambda_k| = 1$ and $\theta_k \in \mathbf{R}$ for $k = 1, 2, \ldots, n$.

17. Use Problem 16 to show that a matrix $U \in M_n$ is unitary and symmetric if and only if there is a unitary matrix $V \in M_n$ such that $U = VV^T$.

18. We have shown that every matrix $A \in M_n$ is similar to a symmetric matrix. Is every matrix similar to a Hermitian matrix? To a normal matrix?

19. Use (4.4.9) to show that every matrix is similar to its transpose.

20. Show that Theorem (4.4.9) is not true over the real field; that is, show that not every matrix $A \in M_n(\mathbf{R})$ is similar to a real symmetric matrix.

21. A complex symmetric matrix A can have an isotropic vector v as an eigenvector; that is, $Av = \lambda v$, $v \neq 0$, and $v^T v = 0$. If A is diagonalizable, however, show that λ cannot be a simple eigenvalue. *Hint:* Write $A = S\Lambda S^{-1}$ with v the first column of S and argue that $S^T S$ is singular because its first row is zero. In particular, if $v \in \mathbf{C}^n$ is any vector such that $v^T v = 0$, the symmetric (rank 1) matrix $A = vv^T$ cannot be diagonalizable. See Problem 15.

22. Provide the details for the following outline for another proof of Corollary (4.4.4). The notation and hypotheses are as in (4.4.4). This is essentially Siegel's (1943) proof. (a) $A\bar{A}$ is Hermitian, so there is a unitary $V \in M_n$ and a real diagonal $\Lambda_1 \in M_n$ such that $A\bar{A} = V\Lambda_1 V^*$. (b) $V^* A \bar{V} \equiv B$ is both symmetric and normal, so by (4.4.7) there is a diagonal $\Lambda \in M_n$ and a real orthogonal matrix $Q \in M_n(\mathbf{R})$ such that $B = Q\Lambda Q^T$. (c) $A = (VQ)\Lambda(VQ)^T$. Now write $\Lambda = E\Sigma E^T$ with E, Σ both diagonal and Σ nonnegative to get $A = U\Sigma U^T$ with $U = VQE$ unitary.

23. Let $z = [z_1, z_2, \ldots, z_n]^T$ be a vector of n complex variables and let $f(z)$ be a complex analytic function of n complex variables in some domain $D \subset \mathbf{C}^n$. Because of the equality of the mixed partial derivatives, $H = [\partial^2 f/\partial z_i\, \partial z_j]$ is symmetric at every point $z \in D$. The discussion in (4.0.3) shows that one may assume that the coefficient matrix $A = [a_{ij}]$ in the general linear partial differential operator

$$Lf = \sum_{i,j=1}^{n} a_{ij}(z) \frac{\partial^2 f}{\partial z_i\, \partial z_j}$$

is symmetric. Show that at each point $z_0 \in D$ there is a unitary change of variables $z \to U\zeta$ such that in the new coordinates Lf is diagonal at z_0, that is,

$$Lf = \sum_{i=1}^{n} \sigma_i \frac{\partial^2 f}{\partial \zeta_i^2}, \qquad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0 \qquad \text{at} \quad z = z_0$$

24. Use (4.4.13) and an induction argument like the one used in the proof of (1.3.19) to prove the following analog of Theorem (4.1.6) on simultaneous unitary diagonalization of a family of Hermitian matrices:

Let $\mathfrak{F} \subset M_n$ be a given family of diagonalizable symmetric matrices. There exists a complex orthogonal matrix Q such that QAQ^T is diagonal for all $A \in \mathfrak{F}$ if and only if $\mathfrak{F}$ is a commuting family.

25. Use the argument in the proof of Theorem (4.4.7) to show that a matrix $A \in M_n$ is both skew-symmetric ($A = -A^T$) and normal if and only if there is a real orthogonal matrix $Q \in M_n(\mathbf{R})$ such that $Q^T A Q = 0 \oplus 0 \oplus \cdots \oplus 0 \oplus A_1 \oplus A_2 \oplus \cdots \oplus A_k$, where each $A_j \in M_2$ has the form

$$A_j = \begin{bmatrix} 0 & z_j \\ -z_j & 0 \end{bmatrix}, \qquad z_j \in \mathbf{C}, \qquad j = 1, 2, \ldots, k \tag{4.4.14}$$

Hint: Consider the real and imaginary parts of A and use Theorem (2.5.15). When are the 1-by-1 zero direct summands absent?

26. Use Problem 25 and the argument in Problem 22 to prove a complex skew-symmetric analog of Takagi's factorization (4.4.4) for a complex symmetric matrix: A matrix $A \in M_n$ is skew-symmetric ($A = -A^T$) if and only if there is a unitary $U \in M_n$ such that

$$A = U(0 \oplus \cdots \oplus 0 \oplus A_1 \oplus \cdots \oplus A_k)U^T$$

where each $A_j \in M_n$ has the form (4.4.14). In particular, conclude that a skew-symmetric complex matrix must have even rank.

27. Let $W \in M_n$ be a given unitary matrix. Show that there is a unitary $V \in M_n$ such that $V^2 = W$ and $V^T A = AV$ whenever $A \in M_n$ is such that $W^T A = AW$. *Hint:* If $W = U\Lambda U^*$ with U unitary and $\Lambda = \operatorname{diag}(e^{i\theta_1}, \ldots, e^{i\theta_n})$ with $0 \leq \theta_j < 2\pi$, consider the natural square root $\Lambda^{1/2} \equiv \operatorname{diag}(e^{i\theta_1/2}, \ldots, e^{i\theta_n/2})$ and let $V \equiv U\Lambda^{1/2}U^*$. Show that $W^T A = AW$ if and only if Λ commutes with $U^T A U$. Use the argument in the proof of (1.3.12), or show that V is a polynomial in W, and conclude that $\Lambda^{1/2}$ commutes with $U^T A U$ and hence $V^T A = AV$.

28. Provide the details for the following outline of yet another proof of Corollary (4.4.4). The notation and hypotheses are as in (4.4.4). This is essentially Hua's (1944) proof. Assume first that A is nonsingular. (a) $A\bar{A}$ is Hermitian and positive definite ($x^* A\bar{A} x = (\bar{A}x)^*(\bar{A}x) \geq 0$ for all $x \in \mathbf{C}^n$), so there is a unitary $Z \in M_n$ and a nonnegative nonsingular diagonal $\Sigma \in M_n$ such that $A\bar{A} = Z\Sigma^2 Z^*$. (b) $W \equiv \Sigma^{-1} Z^* A \bar{Z}$ is unitary and ΣW is symmetric, so $\Sigma W = W^T \Sigma$. (c) Use Problem 27 to show that there is a unitary $V \in M_n$ such that $V^2 = W$ and $\Sigma V = V^T \Sigma$. (d) $Z^* A \bar{Z} = \Sigma W = \Sigma V^2 = (\Sigma V)V = V^T \Sigma V$, so $A = (ZV^T)\Sigma(ZV^T)^T$. Let $U = ZV^T$. (e) If A is singular, employ the argument at the beginning of Problem 2 to reduce to the nonsingular case.

Further Readings and Notes. For the original versions of Corollary (4.4.4) see T. Takagi, "On an Algebraic Problem Related to an Analytic Theorem of Caratheodory and Fejer and on an Allied Theorem of Landau," *Japan. J. Math.* 1 (1925), 83–93, as well as I. Schur, "Ein Satz über Quadratische Formen mit Komplexen Koeffizienten," *Amer. J. Math.* 67 (1945), 472–480. Other proofs were given by C. L. Siegel, "Symplectic Geometry," *Amer. J. Math.* 65 (1943), lemma 1, pp. 12, 14–15; L.-K. Hua, "On the Theory of Automorphic Functions of a Matrix Variable I - Geometric Basis," *Amer. J. Math.* 66 (1944), 470–488; and N. Jacobson, "Normal Semi-Linear Transformations," *Amer. J. Math.* 61 (1939), 45–58. The proof of (4.4.4) via the triangular reduction (4.4.3) is in Y. P. Hong and R. A. Horn, "On the Reduction of a Matrix to Triangular or Diagonal Form by Consimilarity," *SIAM J. Algebraic and Discrete Methods* 7 (1986), 80–88. For generalizations of (4.4.11) to an arbitrary field, see O. Taussky, "The Role of Symmetric Matrices in the Study of General Matrices," *Linear Algebra Appl.* 5 (1972), 147–154.

4.5 Congruence and simultaneous diagonalization of Hermitian and symmetric matrices

Any real second-order linear partial differential operator L can be written in the form

$$Lf=\sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial^2 f(x)}{\partial x_i\,\partial x_j}+\text{lower-order terms},\qquad x=[x_i]\in D\subset \mathbf{R}^n \tag{4.5.1}$$

where we assume that the coefficients $a_{ij}(x)$ are defined on some domain $D\subset\mathbf{R}^n$ and that the function f is twice continuously differentiable on D. As we saw in (4.0.3), we may assume without loss of generality that the matrix of coefficients $A(x)=[a_{ij}(x)]$ is a real symmetric matrix for all $x\in D$. By "lower-order terms" we mean terms involving f and its first partial derivatives only.

If we make a nonsingular change of independent variables to new variables $s=[s_i]\in D\subset\mathbf{R}^n$, then each $s_i=s_i[x]=s_i(x_1,\ldots,x_n)$, and nonsingularity means that the Jacobian matrix

$$S(x)=\left[\frac{\partial s_i(x)}{\partial x_j}\right]\in M_n$$

is nonsingular at each point of D. This assumption guarantees that the inverse change of variables $x=x(s)$ exists locally. It is a straightforward application of the chain rule to show that, in these new coordinates, the operator L has the form

$$\begin{aligned}Lf&=\sum_{i,j=1}^{n}\left[\sum_{p,q=1}^{n}\frac{\partial s_i}{\partial x_p}a_{pq}\frac{\partial s_j}{\partial x_q}\right]\frac{\partial^2 f}{\partial s_i\,\partial s_j}+\text{lower-order terms}\\&=\sum_{i,j=1}^{n} b_{ij}\frac{\partial^2 f}{\partial s_i\,\partial s_j}+\text{lower-order terms}\end{aligned} \tag{4.5.2}$$

Thus, the new matrix of coefficients B (in the coordinates $s=[s_i]$) is related to the old matrix of real coefficients A (in the coordinates $x=[x_i]$) by the relation

$$B=SAS^T \tag{4.5.3^T}$$

where S is a real nonsingular matrix.

If the differential operator L is associated with some physical law (e.g., the Laplacian $L=\nabla^2$ and electrostatic potentials), the choice of coordinates for the independent variable should not affect the law, although it obviously affects the form of L. Thus, we are led to ask what the invariants are of the set of all matrices B that are related to a given matrix A by the relation (4.5.3^T).

Another example of a transformation like (4.5.3^T) comes from probability and statistics. Suppose that $X_1,X_2,\ldots,X_n$ are real or complex random variables with finite second moments on some probability space with expectation operator E, and let $\mu_i=E(X_i)$ denote the respective means. The Hermitian matrix $A=[a_{ij}]=(E[(X_i-\mu_i)(\bar X_j-\bar\mu_j)])\equiv \operatorname{Cov}(X)$ is the *covariance matrix* of the random vector $X=[X_1,\ldots,X_n]^T$. If $S=[s_{ij}]\in M_n$ is a given matrix, SX is a random vector whose components are linear combinations of the components of X. The means of the components of SX are

$$E((SX)_i)=E\left(\sum_{k=1}^{n}s_{ik}X_k\right)=\sum_{k=1}^{n}s_{ik}E(X_k)=\sum_{k=1}^{n}s_{ik}\mu_k$$

and the covariance matrix of SX is

$$\begin{aligned}\operatorname{Cov}(SX)&=(E[((SX)_i-E((SX)_i))((\overline{SX})_j-E((\overline{SX})_j))])\\&=\left(E\left[\left(\sum_{p=1}^{n}s_{ip}(X_p-\mu_p)\right)\left(\sum_{q=1}^{n}\bar s_{jq}(\bar X_q-\bar\mu_q)\right)\right]\right)\\&=\left(\sum_{p,q=1}^{n}s_{ip}E[(X_p-\mu_p)(\bar X_q-\bar\mu_q)]\bar s_{jq}\right)=\left(\sum_{p,q=1}^{n}s_{ip}a_{pq}\bar s_{iq}\right)\\&=SAS^*\end{aligned}$$

This shows that

$$\operatorname{Cov}(SX)=S\operatorname{Cov}(X)S^* \tag{4.5.3*}$$

Thus, the covariance matrix of a random vector transforms according to a slightly different law from ($4.5.3^T$), but it reduces to ($4.5.3^T$) if the matrix S is real.

As a final example, consider the general quadratic form

$$Q_A(x) = \sum_{i,j=1}^{n} a_{ij} x_i x_j = x^T A x, \qquad x = [x_i] \in \mathbf{C}^n$$

and the Hermitian form

$$H_B(x) = \sum_{i,j=1}^{n} b_{ij} \bar{x}_i x_j = x^* B x, \qquad x = [x_i] \in \mathbf{C}^n$$

where $A = [a_{ij}]$ and $B = [b_{ij}]$. If $S \in M_n$ is a given matrix, then

$$Q_A(Sx) = (Sx)^T A(Sx) = x^T (S^T A S) x = Q_{S^T AS}(x)$$

$$H_A(Sx) = (Sx)^* B(Sx) = x^* (S^* B S) x = H_{S^* BS}(x)$$

In this example it does not matter whether A, B, S, and x are real or complex. There are two slightly different laws of transformation at work here, and this is the reason for the following definition.

4.5.4 **Definition.** Let $A, B \in M_n$ be given. If there exists a nonsingular matrix S such that

(a) $B = SAS^*$, then B is said to be **congruent* ("star-congruent") to A.

(b) $B = SAS^T$, then B is said to be *^{T}congruent* ("tee-congruent") to A.

It is clear that these two notions of congruence must be closely related; they are the same if S is a real matrix. When it is not important to distinguish between the two notions, we use the term *congruence* without a prefix. Some authors use the term *conjunctive* for *congruent, but we have chosen to deviate from this usage to have terms with greater mnemonic content.

Exercise. Show that congruent matrices have the same rank.

Notice that if A is Hermitian, then so is SAS^* (even if S is singular); if A is symmetric, then SAS^T is also symmetric. Usually, one is interested in congruences that preserve the type of the matrix, *congruence for Hermitian matrices and ^{T}congruence for symmetric matrices. If A is real and symmetric, however, then it is both symmetric and Hermitian; SAS^* is then Hermitian and SAS^T is symmetric. For a real symmetric matrix, one

may wish to consider either *congruence or ^{T}congruence, depending on the context. Both types of congruence share an important property with similarity.

4.5.5 **Theorem.** Both *congruence and ^{T}congruence are equivalence relations. That is, for any $A \in M_n$,

(a) A is congruent to A.

(b) If A is congruent to B, then B is congruent to A.

(c) If A is congruent to B and B is congruent to C, then A is congruent to C.

Proof: For (a), write $A = IAI^*$. If $A = SBS^*$ and S is nonsingular, then $B = S^{-1}A(S^{-1})^*$. Finally, if $A = S_1 B S_1^*$ and $B = S_2 C S_2^*$, then $A = (S_1 S_2) C (S_1 S_2)^*$. The proof for ^{T}congruence is formally the same. □

The set of all n-by-n matrices is therefore partitioned into equivalence classes by congruence. As an abstract problem, we may seek a canonical representative of each equivalence class under each type of congruence. This problem is more complicated for *congruence, so we take up this case first.

The practical problem of understanding and classifying differential operators by identifying the invariants of the congruence relation leads us to the problem of identifying a canonical representative of the equivalence class of real symmetric matrices that are congruent (via a real matrix S) to a given matrix. It turns out that this problem has a simple solution: Just count the number of positive, negative, and zero eigenvalues. For this reason, we introduce the following terminology:

4.5.6 **Definition.** Let $A \in M_n$ be a Hermitian matrix. The *inertia* of A is the ordered triple

$$i(A) = (i_+(A), i_-(A), i_0(A))$$

where $i_+(A)$ is the number of positive eigenvalues of A, $i_-(A)$ is the number of negative eigenvalues of A, and $i_0(A)$ is the number of zero eigenvalues of A, all counting multiplicity. Notice that the rank of A is equal to $i_+(A) + i_-(A)$. The *signature* of A is the quantity $i_+(A) - i_-(A)$.

Exercise. Show that the inertia of a Hermitian matrix $A \in M_n$ is uniquely determined if one knows both the signature and the rank of A, and conversely.

If $A \in M_n$ is a given Hermitian matrix, then $A = U\Lambda U^*$ with $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ and U unitary. It is convenient to assume that the positive eigenvalues occur first among the diagonal entries of Λ, then the negative eigenvalues, and then the zero eigenvalues. Thus $\lambda_1, \lambda_2, \ldots, \lambda_{i_+} > 0$, $\lambda_{i_+ +1}, \ldots, \lambda_{i_+ + i_-} < 0$, and $\lambda_{i_+ + i_- +1} = \cdots = \lambda_n = 0$. If we set

$$D = \operatorname{diag}(+\sqrt{\lambda_1}, \ldots, +\sqrt{\lambda_{i+}}, +\sqrt{-\lambda_{i_+ +1}}, \ldots, +\sqrt{-\lambda_{i_+ + i_-}}, 1, \ldots, 1)$$

then D is a nonsingular real diagonal matrix, and

$$\Lambda = D \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & 0 & \\ & & 1 & & & & & & \\ & & & -1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & -1 & & & \\ & & & & & & 0 & & \\ & 0 & & & & & & \ddots & \\ & & & & & & & & 0 \end{bmatrix} D$$

where the indicated matrix has exactly $i_+(A)$ "+1" terms, $i_-(A)$ "−1" terms, and $i_0(A)$ "0" terms. Thus, the matrix A can be written as

$$A = U\Lambda U^* = S \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & 0 & \\ & & 1 & & & & & & \\ & & & -1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & -1 & & & \\ & & & & & & 0 & & \\ & 0 & & & & & & \ddots & \\ & & & & & & & & 0 \end{bmatrix} S^* = SI(A)S^* \tag{4.5.7}$$

where $S \equiv UD$ is a nonsingular matrix, and $I(A)$ is the *inertia matrix* of A. Thus, every Hermitian matrix is *congruent to a diagonal matrix of very simple form that is known once the inertia of the matrix is known. It would be attractive to use the inertia matrix as the canonical representative of the equivalence class of matrices that are *congruent to A, but in order to do so we must be certain that *congruent Hermitian matrices have the same inertia. This is the content of the following theorem, which is usually known as *Sylvester's law of inertia.*

4.5.8 **Theorem.** Let $A, B \in M_n$ be Hermitian matrices. There is a nonsingular matrix $S \in M_n$ such that $A = SBS^*$ if and only if A and B have the same inertia, that is, the same number of positive, negative, and zero eigenvalues.

Proof: If A and B have the same inertia, then each can be represented in the form (4.5.7) with a possibly different S for each, but with the same inertia matrix. Since the *congruence relation is transitive, and since A and B are *congruent to the same matrix, they are *congruent to each other. It is the converse that is more interesting.

Suppose A and B are *congruent and that $A = SBS^*$ for some nonsingular matrix $S \in M_n$. Since congruent matrices have the same rank, $i_0(A) = i_0(B)$ and we need only show that $i_+(A) = i_+(B)$. Let $v_1, v_2, \ldots, v_{i_+(A)}$ be orthonormal eigenvectors of A corresponding to the positive eigenvalues $\lambda_1(A), \ldots, \lambda_{i_+(A)}(A)$, and let $S_+(A) = \operatorname{Span}\{v_1, \ldots, v_{i_+(A)}\}$. The dimension of $S_+(A)$ is $i_+(A)$, and if $x = \alpha_1 v_1 + \cdots + \alpha_{i_+(A)} v_{i_+(A)} \neq 0$, then $x^*Ax = \lambda_1(A)|\alpha_1|^2 + \cdots + \lambda_{i_+(A)}(A)|\alpha_{i_+(A)}|^2 > 0$. But then

$$x^*SBS^*x = (S^*x)^*B(S^*x) > 0$$

so $y^*By > 0$ for all nonzero vectors y in $\operatorname{Span}\{S^*v_1, \ldots, S^*v_{i_+(A)}\}$, which has dimension $i_+(A)$. By (4.3.23), we must have $i_+(B) \geq i_+(A)$. But since the roles of A and B in this argument can be reversed, we conclude that $i_+(B) = i_+(A)$. □

Exercise. Let $A \in M_n$ be Hermitian. Show that A is *congruent to the identity matrix if and only if all the eigenvalues of A are positive.

Exercise. Let $A, B \in M_n$ be real symmetric matrices. Show that A and B are *congruent via a complex matrix if and only if they are congruent via a real matrix.

Exercise. Let $A, B \in M_n$ be real symmetric matrices. Show that A and B are congruent via a *real* matrix if and only if A and B have the same inertia.

Exercise. How many disjoint equivalence classes under *congruence are there in the set of n-by-n complex Hermitian matrices? In the set of n-by-n real symmetric matrices?

Sylvester's theorem settles completely the question of choosing a representative element from each eigenvalue class of Hermitian matrices under *congruence by guaranteeing that the *signs* of the eigenvalues of a Hermitian matrix do not change under *congruence. But how do the *magnitudes* of the eigenvalues change under *congruence? Using the

simplest form of Weyl's theorem (4.3.1), we can give a quantitative form of Sylvester's theorem.

4.5.9 **Theorem** (Ostrowski). Let $A, S \in M_n$ with A Hermitian and S nonsingular. Let the eigenvalues of A and SS^* be arranged in increasing order (4.2.1). For each $k=1,2,\ldots$, there exists a positive real number θ_k such that $\lambda_1(SS^*) \le \theta_k \le \lambda_n(SS^*)$ and

$$\lambda_k(SAS^*) = \theta_k \lambda_k(A) \tag{4.5.10}$$

Proof: First observe that if $SS^*x = \lambda x$ and $x \ne 0$, then $\lambda = x^*SS^*x/x^*x = (S^*x)^*(S^*x)/x^*x > 0$, so all the eigenvalues of SS^* are positive. Let k be a given integer, $1 \le k \le n$, and consider the Hermitian matrix $A - \lambda_k(A)I$, whose kth eigenvalue is zero. By Sylvester's theorem (4.5.8), the kth eigenvalue of $S(A-\lambda_k(A)I)S^* = SAS^* - \lambda_k(A)SS^*$ is also zero. Weyl's inequality (4.3.2) says that the kth eigenvalue of $SAS^* - \lambda_k(A)SS^*$ has the bounds

$$\begin{aligned}\lambda_k(SAS^*) + \lambda_1(-\lambda_k(A)SS^*) &\le \lambda_k(SAS^* - \lambda_k(A)SS^*) = 0\\ &\le \lambda_k(SAS^*) + \lambda_n(-\lambda_k(A)SS^*)\end{aligned}$$

or that

$$\begin{aligned}\lambda_k(SAS^*) &\le -\lambda_1(-\lambda_k(A)SS^*) = \lambda_n(\lambda_k(A)SS^*)\\ &= \begin{cases}\lambda_k(A)\lambda_n(SS^*) & \text{if} \quad \lambda_k(A) \ge 0\\ \lambda_k(A)\lambda_1(SS^*) & \text{if} \quad \lambda_k(A) \le 0\end{cases}\end{aligned}$$

and

$$\begin{aligned}\lambda_k(SAS^*) &\ge -\lambda_n(-\lambda_k(A)SS^*) = \lambda_1(\lambda_k(A)SS^*)\\ &= \begin{cases}\lambda_k(A)\lambda_1(SS^*) & \text{if} \quad \lambda_k(A) \ge 0\\ \lambda_k(A)\lambda_n(SS^*) & \text{if} \quad \lambda_k(A) \le 0\end{cases}\end{aligned}$$

In either case [$\lambda_k(A) \ge 0$ or $\lambda_k(A) \le 0$], these inequalities imply that $\lambda_k(SAS^*) = \theta_k \lambda_k(A)$ for some θ_k with $\lambda_1(SS^*) \le \theta_k \le \lambda_n(SS^*)$. □

If $A = I \in M_n$ in Ostrowski's theorem, then all $\lambda_k(A) = 1$ and $\theta_k = \lambda_k(SS^*)$. If $S \in M_n$ is unitary, then $\lambda_1(SS^*) = \lambda_n(SS^*) = 1$ and all $\theta_k = 1$; this expresses the invariance of the eigenvalues under a unitary similarity. Thus, the bounds for θ_k given in the theorem are best possible for any given A as well as for any given nonsingular S.

By a simple continuity argument, Ostrowski's theorem can be extended to cover the situation in which S is singular. In this event, let $\epsilon > 0$ and apply the theorem with S replaced by $S + \epsilon I$ to see that $\lambda_k((S+\epsilon I)A(S+\epsilon I)^*) = \theta_k \lambda_k(A)$ with $\lambda_1((S+\epsilon I)(S+\epsilon I)^*) \le \theta_k \le \lambda_n((S+\epsilon I)(S+\epsilon I)^*)$. Now let

$\epsilon \to 0$ to obtain the bound $0 \le \theta_k \le \lambda_n(SS^*)$. This result may be thought of as an extension of Sylvester's law of inertia to singular *congruences.

4.5.11 **Corollary.** Let $A, S \in M_n$ and let A be Hermitian. Let the eigenvalues of A and SS^* be arranged in increasing order (4.2.1). For each $k = 1, 2, \ldots, n$ there exists a nonnegative real number θ_k such that $\lambda_1(SS^*) \le \theta_k \le \lambda_n(SS^*)$ and

$$\lambda_k(SAS^*) = \theta_k \lambda_k(A)$$

In particular, the number of positive (negative) eigenvalues of SAS^* is less than or equal to the number of positive (negative) eigenvalues of A.

The problem of finding a canonical representative of the equivalence classes of complex symmetric matrices under ^{T}congruence has an even simpler solution: Just compute the rank.

4.5.12 **Theorem.** Let $A, B \in M_n$ be (complex or real) symmetric matrices. There is a nonsingular matrix $S \in M_n$ such that $A = SBS^T$ if and only if A and B have the same rank.

Proof: If $A = SBS^T$ with S nonsingular, then A has the same rank as B by (0.4.6). Conversely, use (4.4.4) to write

$$A = U_1 \Sigma_1 U_1^T = U_1 I(\Sigma_1) D_1^2 U_1^T = (U_1 D_1) I(\Sigma_1)(U_1 D_1)^T$$

where $I(\Sigma_1)$ is the inertia matrix (4.5.7) of Σ_1 and is determined solely by the rank of A, U_1 is unitary, $\Sigma_1 = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ with all $\sigma_i \ge 0$, and $D_1 = \operatorname{diag}(d_1, d_2, \ldots, d_n)$ with

$$d_i = \begin{cases}\sqrt{\sigma_i} & \text{if} \quad \sigma_i > 0\\ 1 & \text{if} \quad \sigma_i = 0\end{cases}$$

Notice that D_1 is nonsingular. In the same way, we can also write $B = (U_2 D_2) I(\Sigma_2)(U_2 D_2)^T$ with similar definitions. If we assume that rank A = rank B, then $I(\Sigma_1) = I(\Sigma_2)$, and

$$I(\Sigma_1) = (U_1 D_1)^{-1} A[(U_1 D_1)^T]^{-1} = I(\Sigma_2) = (U_2 D_2)^{-1} B[(U_2 D_2)^T]^{-1}$$

and hence

$$A = (U_1 D_1)(U_2 D_2)^{-1} B[(U_1 D_1)(U_2 D_2)^{-1}]^T$$

We conclude that A and B are ^{T}congruent. □

Exercise. How many disjoint equivalence classes under ^{T}congruence are there in the set of n-by-n complex symmetric matrices? In the set of n-by-n real symmetric matrices?

Exercise. Let $A \in M_n$ be symmetric. Show that there exists a nonsingular $S \in M_n$ such that $A = SS^T$ if and only if A is nonsingular.

Exercise. Let $A, B \in M_n$ be symmetric. Show that there exist two nonsingular matrices $X, Y \in M_n$ such that $A = XBY$, that is, A and B are equivalent, if and only if there exists a nonsingular $S \in M_n$ such that $A = SBS^T$, that is, A and B are ^{T}congruent. *Hint:* If $A = XBY$, how are the ranks of A and B related?

The preceding result is an analog of Sylvester's law of inertia (4.5.8) for ^{T}congruence of complex matrices. The following result is an analog of Ostrowski's quantitative version [(4.5.9) and (4.5.11)] of Sylvester's theorem.

4.5.13 **Theorem.** Let $A, S \in M_n$ with $A = A^T$. Let $A = U\Sigma U^T$ and $SAS^T = VMV^T$ be Takagi factorizations (4.4.4) of A and SAS^T with U and V unitary, $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$, and $M = \text{diag}(\mu_1, \mu_2, \ldots, \mu_n)$ with all $\sigma_i, \mu_i \geq 0$. Let $\lambda_i(SS^*)$ denote the eigenvalues of SS^*. Suppose the numbers σ_i, μ_i, and $\lambda_i(SS^*)$ are all arranged in increasing order (4.2.1). For each $k = 1, 2, \ldots, n$ there exists a nonnegative real number θ_k with $\lambda_1(SS^*) \leq \theta_k \leq \lambda_n(SS^*)$ such that $\mu_k = \theta_k \sigma_k$. If S is nonsingular, all $\theta_k > 0$.

Proof: The numbers μ_i^2 are the eigenvalues of BB^*, where $B = SAS^T$. Thus,

$$\mu_k^2 = \lambda_k(BB^*) = \lambda_k(SAS^T\bar{S}\bar{A}S^*) = \lambda_k(S[AS^T\bar{S}\bar{A}]S^*) = \hat{\theta}_k \lambda_k(AS^T\bar{S}\bar{A})$$

for some $\hat{\theta}_k$ with $\lambda_1(SS^*) \leq \hat{\theta}_k \leq \lambda_n(SS^*)$; we have used (4.5.11) to obtain the last equality. Since the eigenvalues of a product of two matrices are independent of the order of the product (1.3.20), we also have

$$\mu_k^2 = \hat{\theta}_k \lambda_k(AS^T\bar{S}\bar{A}) = \hat{\theta}_k \lambda_k(\bar{S}\bar{A}AS^T) = \hat{\theta}_k \lambda_k(SA\bar{A}S^*)$$

because the eigenvalues λ_k are real. Applying (4.5.11) again, we obtain

$$\mu_k^2 = \hat{\theta}_k \tilde{\theta}_k \lambda_k(A\bar{A}) = \hat{\theta}_k \tilde{\theta}_k \sigma_k^2$$

for some $\tilde{\theta}_k$ with $\lambda_1(SS^*) \leq \tilde{\theta}_k \leq \lambda_n(SS^*)$. Thus, $\mu_k = \sqrt{\hat{\theta}_k \tilde{\theta}_k}\, \sigma_k = \theta_k \sigma_k$ and $\theta_k \equiv \sqrt{\hat{\theta}_k \tilde{\theta}_k}$ satisfies the required bounds. □

We know from (1.3.19) that two diagonalizable matrices can be simultaneously diagonalized by the same similarity if and only if they commute. What is the corresponding result for simultaneous diagonalization by congruence?

Perhaps the earliest motivation for results about simultaneous diagonalization by congruence came from mechanics in the study of "small oscillations" about a stable equilibrium. If the configuration of a dynamical system is specified by generalized (Lagrangian) coordinates $q_1, q_2, \ldots, q_n$ in which the origin is a point of stable equilibrium, then near the origin the potential energy function V can be approximated by a real quadratic form

$$V = \sum_{i,j=1}^{n} a_{ij} q_i q_j$$

in the generalized coordinates q_i. The kinetic energy T can be approximated by a real quadratic form

$$T = \sum_{i,j=1}^{n} b_{ij} \dot{q}_i \dot{q}_j$$

in the generalized velocities $\dot{q}_i$. The behavior of the system is governed by Lagrange's equations

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_i}\right) - \frac{\partial T}{\partial q_i} + \frac{\partial V}{\partial q_i} = 0$$

which are a system of second-order linear ordinary differential equations with constant coefficients. These equations are *coupled* (and hence more difficult to solve) if the two quadratic forms T and V are not diagonal. We may assume that the real matrices $A = [a_{ij}]$ and $B = [b_{ij}]$ are symmetric.

If a real nonsingular transformation $S = [s_{ij}] \in M_n$ can be found such that SAS^T and SBS^T are both diagonal, then with respect to new generalized coordinates p_i with

$$q_i = \sum_{j=1}^{n} s_{ij} p_j \tag{4.5.14}$$

the kinetic energy and potential energy quadratic forms T and V will both be diagonal. In this event, Lagrange's equations will be an *uncoupled* set of n separate second-order linear ordinary differential equations with constant coefficients. These equations can be solved easily in terms of exponentials and trigonometric functions, and the solution to the original problem can be obtained by using (4.5.14).

Thus, a substantial simplification in an important class of mechanics problems can be achieved if we can simultaneously diagonalize two real symmetric matrices by congruence. On physical grounds, the kinetic energy quadratic form is positive definite, and it turns out that this is a sufficient (but not necessary) condition for simultaneous diagonalization by congruence.

There are several types of simultaneous diagonalization results that we might consider. We might have two Hermitian matrices A and B and we

might wish to have UAU^* and UBU^* diagonal for some unitary matrix U, or we might be satisfied with the weaker result of having SAS^* and SBS^* diagonal for some nonsingular matrix S. Similarly, if A and B are symmetric, we might want UAU^T and UBU^T or SAS^T and SBS^T to be diagonal. We might even have a mixed problem with A Hermitian and B symmetric and want UAU^* and UBU^T or SAS^* and SBS^T to be diagonal. In each case, the natural congruence to consider is one that preserves the special algebraic character of the respective matrix. All of these situations arise in the applications. Fortunately, they can all be treated with the same techniques. The simplest case to consider is when one of the two matrices is nonsingular. We present the results in Table 4.5.15T, which gives a list of equivalent necessary and sufficient conditions for each case. The necessary and sufficient conditions are ordered and numbered to suggest parallelism among the various cases.

4.5.15 **Theorem.** Let $A, B \in M_n$ be given. Let U denote a unitary matrix and S a nonsingular matrix with $U, S \in M_n$. Then we have Table 4.5.15T.

Proof: Within each of the six groups of conditions, the equivalence of most of the stated conditions is a matter of definition. The equivalence of (3) and (4) in group I(a) follows from the observation that, if A and B are Hermitian, then AB is Hermitian if and only if A commutes with B, and A is Hermitian if and only if A^{-1} is Hermitian. The equivalence of III(a)(3) and (4) is established similarly, since B is symmetric if and only if B^{-1} is symmetric, and $A^T = \bar{A}$ if A is Hermitian.

Within each of the six groups, the necessity of condition (1) follows directly from the assumption that the respective congruences are in diagonal form. For example, in case II(b) if $SAS^T = \Lambda$ and $SBS^T = M$ are both diagonal, then

$$A^{-1}B = (S^T\Lambda^{-1}S)[S^{-1}M(S^T)^{-1}] = S^T(\Lambda^{-1}M)(S^T)^{-1}$$

so $R = S^T$ will diagonalize $C = A^{-1}B$. Similarly, in cases I(b) and III(b), $R = S^*$ will work. If S is unitary, the corresponding matrix R in each case is also unitary.

Consider the cases I in which A and B are Hermitian and A is nonsingular. Make the assumption I(b)(1) that there is a nonsingular $R = [r_1\, r_2 \cdots r_n] \in M_n$, each $r_i \in \mathbf{C}^n$, and a diagonal $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ with all λ_i real such that $R^{-1}A^{-1}BR = \Lambda$, and hence $BR = AR\Lambda$ and $R^*BR = R^*AR\Lambda$. There is no loss of generality to assume that multiple values of the λ_i terms are grouped together, so that Λ has the block form

Table 4.5.15T.

Assumptions on A and B	That which is to be diagonal	Equivalent necessary and sufficient conditions for simultaneous diagonalization
I. $A = A^*$ $B = B^*$ A nonsingular $C = A^{-1}B$	(a) UAU^* and UBU^*	(1) There is a unitary $V \in M_n$ such that V^*CV is a real diagonal matrix. (2) C has real eigenvalues and is unitarily diagonalizable. (3) C is Hermitian. (4) $AB = BA$, i.e., A commutes with B.
	(b) SAS^* and SBS^*	(1) There is a nonsingular $R \in M_n$ such that $R^{-1}CR$ is real diagonal. (2) C has real eigenvalues and is diagonalizable.
II. $A = A^T$ $B = B^T$ A nonsingular $C = A^{-1}B$	(a) UAU^T and UBU^T	(1) There is a unitary $V \in M_n$ such that V^*CV is diagonal. (2) C is unitarily diagonalizable. (3) C is normal.
	(b) SAS^T and SBS^T	(1) There is a nonsingular $R \in M_n$ such that $R^{-1}CR$ is diagonal. (2) C is diagonalizable.
III. $A = A^*$ $B = B^T$ If A is nonsingular, set $C = A^{-1}B$. If B is nonsingular, set $C = B^{-1}A$.	(a) UAU^* and UBU^T	(1) There is a unitary $W \in M_n$ such that $W^{-1}C\bar{W}$ is diagonal. (3) C is symmetric. (4) $AB = B\bar{A}$.
	(b) SAS^* and SBS^T	(1) There is a nonsingular $R \in M_n$ such that $R^{-1}C\bar{R}$ is diagonal. (5) There is a nonsingular $R \in M_n$ such that $R^{-1}C\bar{R}$ is symmetric.

$$\Lambda = \begin{bmatrix} \Lambda_1 & & & 0 \\ & \Lambda_2 & & \\ & & \ddots & \\ 0 & & & \Lambda_k \end{bmatrix} \tag{4.5.16}$$

$$\Lambda_i \in M_{n_i}, \quad 1 \le n_i \le n; \quad \Lambda_i = \mu_i I, \quad i = 1, 2, \ldots, k$$

with all μ_i real and $\mu_i \neq \mu_j$ if $i \neq j$. If not all the λ_i terms are equal, choose any i, j with $1 \le i, j \le n$ such that $\lambda_i \neq \lambda_j$ and consider the i, j entry of both sides of the identity $R^*BR = R^*AR\Lambda$. This is

$$r_i^*Ar_j\lambda_j = r_i^*Br_j = \overline{r_j^*Br_i} = \overline{r_j^*Ar_i\lambda_i} = r_i^*Ar_j\lambda_i$$

where we have used the facts that A and B are Hermitian (so $x^*Ay = \overline{y^*Ax}$ for all $x, y \in \mathbf{C}^n$) and that λ_i and λ_j are real. Since $\lambda_i \neq \lambda_j$, we conclude that $r_i^*Ar_j = 0$ and hence that $r_j^*Ar_i = r_i^*Br_j = r_j^*Br_i = 0$. This means that the matrices R^*BR and R^*AR are both block diagonal and conformal with (4.5.16); that is,

$$R^*BR = \begin{bmatrix} B_1 & & & 0 \\ & B_2 & & \\ & & \ddots & \\ 0 & & & B_k \end{bmatrix} = R^*AR\Lambda$$

$$= \begin{bmatrix} \mu_1 A_1 & & & 0 \\ & \mu_2 A_2 & & \\ & & \ddots & \\ 0 & & & \mu_k A_k \end{bmatrix}$$

with $B_i, A_i \in M_{n_i}$ for $i = 1, 2, \ldots, k$. This is a partial reduction to diagonal form that is complete if $k = n$, that is, if all λ_i are distinct. If $k < n$, then some block has $n_i > 1$ and $B_i = \mu_i A_i$. Since A_i and B_i are Hermitian, we may use the spectral theorem (4.1.5) to write $A_i = U_i D_i U_i^*$ with $U_i, D_i \in M_{n_i}$, U_i unitary, and D_i real diagonal. Then $B_i = \mu_i A_i = U_i(\mu_i D_i)U_i^*$ is diagonalized as well. If we set

$$U = \begin{bmatrix} U_1 & & & 0 \\ & U_2 & & \\ & & \ddots & \\ 0 & & & U_k \end{bmatrix}, \quad D = \begin{bmatrix} D_1 & & & 0 \\ & D_2 & & \\ & & \ddots & \\ 0 & & & D_k \end{bmatrix}$$

with $U_i = [1]$ if $n_i = 1$, then U is unitary, D is real diagonal, and

$$R^*BR = U(D\Lambda)U^* \quad \text{and} \quad R^*AR = UDU^*$$

Finally, we have the required representations

$$A = [(R^{-1})^*U]D[(R^{-1})^*U]^* \quad \text{and} \quad B = [(R^{-1})^*U](D\Lambda)[R^{-1})^*U]^*$$

Notice that if we assume I(a)(1), then the argument is the same except that we also know that the matrix R is unitary. In this event, $(R^{-1})^*U = RU$ is unitary and the sufficiency of I(a)(1) is proved.

The arguments required in the remaining four cases are similar. One uses the respective hypothesis to obtain congruent matrices that are block diagonal and then uses the spectral theorem for Hermitian matrices or Takagi's factorization (4.4.4) for symmetric matrices to complete the reduction to diagonal form.

Consider the cases II in which A and B are symmetric and A is nonsingular. Make the assumption II(b)(1) that there is a nonsingular $R = [r_1\, r_2 \cdots r_n] \in M_n$, each $r_i \in \mathbf{C}^n$, and a diagonal $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ (not necessarily real) such that $R^{-1}A^{-1}BR = \Lambda$, and hence $BR = AR\Lambda$ and $R^TBR = R^TAR\Lambda$. Assume again that the multiple values of the λ_i terms are grouped together so that Λ has the form (4.5.16) with all μ_i distinct. If not all the λ_i terms are equal, choose any i, j with $1 \le i, j \le n$ such that $\lambda_i \neq \lambda_j$ and consider the i, j entry of both sides of the identity $R^TBR = R^TAR\Lambda$. This is

$$r_i^TAr_j\lambda_j = r_i^TBr_j = r_j^TBr_i = r_j^TAr_i\lambda_i = r_i^TAr_j\lambda_i$$

where we have used the symmetry of A and B ($x^TAy = y^TAx$ for all $x, y \in \mathbf{C}^n$). Since $\lambda_i \neq \lambda_j$, we conclude that $r_i^TAr_j = 0$ and hence that $r_j^TAr_i = r_i^TBr_j = r_j^TBr_i = 0$. This means that the matrices R^TBR and R^TAR are both block diagonal and conformal with (4.5.16); that is,

$$R^TBR = \begin{bmatrix} B_1 & & & 0 \\ & B_2 & & \\ & & \ddots & \\ 0 & & & B_k \end{bmatrix} = R^TAR\Lambda$$

$$= \begin{bmatrix} \mu_1 A_1 & & & 0 \\ & \mu_2 A_2 & & \\ & & \ddots & \\ 0 & & & \mu_k A_k \end{bmatrix}$$

with $B_i, A_i \in M_{n_i}$. If $k = n$, this is the required reduction. If $k < n$, then some block has $n_i > 1$ and $B_i = \mu_i A_i$. Since A_i and B_i are symmetric, we

may use Takagi's factorization (4.4.4) to write $A_i = U_i \Sigma_i U_i^T$ with $U_i, \Sigma_i \in M_{n_i}$, U_i unitary, and Σ_i diagonal with nonnegative diagonal entries. Then $B_i = \mu_i A_i = U_i(\mu_i \Sigma_i)U_i^T$. If we set

$$U = \begin{bmatrix} U_1 & & & 0 \\ & U_2 & & \\ & & \ddots & \\ 0 & & & U_k \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \Sigma_1 & & & 0 \\ & \Sigma_2 & & \\ & & \ddots & \\ 0 & & & \Sigma_k \end{bmatrix}$$

with $U_i = [1]$ if $n_i = 1$, then U is unitary, Σ is diagonal (with nonnegative diagonal entries), and

$$R^T B R = U(\Sigma\Lambda)U^T \quad \text{and} \quad R^T A R = U\Sigma U^T$$

Finally, we have the required representations

$$A = [(R^{-1})^T U]\Sigma[(R^{-1})^T U]^T \quad \text{and} \quad B = [(R^{-1})^T U]\Sigma\Lambda[(R^{-1})^T U]^T$$

If we assume II(a)(1), then R is unitary and $(R^{-1})^T U = \bar{R}U$ is unitary, so the sufficiency of II(a)(1) is also proved.

In the cases III, a slight change in the argument is necessary. Let $A, B \in M_n$ with A Hermitian and nonsingular and B symmetric. Make the assumption III(b)(1) that there is a nonsingular $R = [r_1\, r_2 \cdots r_n] \in M_n$ and a diagonal $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ such that $R^{-1}A^{-1}B\bar{R} = \Lambda$, and hence $B\bar{R} = AR\Lambda$ and $R^*B\bar{R} = \bar{R}^T B\bar{R} = R^*AR\Lambda$. Assume now that the λ_i terms of equal *modulus* are grouped together so that Λ has the form

$$\Lambda = \begin{bmatrix} \Lambda_1 & & & 0 \\ & \Lambda_2 & & \\ & & \ddots & \\ 0 & & & \Lambda_k \end{bmatrix}$$

$$\text{where} \quad \Lambda_i = \begin{bmatrix} \mu_i^{(1)} & & & 0 \\ & \mu_i^{(2)} & & \\ & & \ddots & \\ 0 & & & \mu_i^{(n_i)} \end{bmatrix} \quad \text{for} \quad i = 1, \ldots, k$$

with $|\mu_i^{(j)}| = |\mu_i^{(k)}|$ for $j, k = 1, 2, \ldots, n_i$ and $|\mu_i^{(p)}| \neq |\mu_j^{(q)}|$ if $i \neq j$. If not all the λ_i terms have the same modulus, choose any i, j with $1 \le i, j \le n$ and $|\lambda_i| \neq |\lambda_j|$ and consider the i, j entry of both sides of the identity $\bar{R}^T B\bar{R} = R^*AR\Lambda$. This is

$$r_i^* A r_j \lambda_j = \bar{r}_i^T B \bar{r}_j = \bar{r}_j^T B \bar{r}_i = r_j^* A r_i \lambda_i = \overline{r_i^* A r_j}\, \lambda_i$$

where we have used the fact that A is Hermitian and B is symmetric. From this it follows that $|r_i^*Ar_j|\,|\lambda_j| = |r_i^*Ar_j|\,|\lambda_i|$, and since $|\lambda_i| \neq |\lambda_j|$ we conclude that $r_i^*Ar_j = 0$, and hence that $r_j^*Ar_i = \bar{r}_i^T B\bar{r}_j = \bar{r}_j^T B\bar{r}_i = 0$. This means that the matrices $\bar{R}^T B\bar{R}$ and R^*AR are both block diagonal and conformal with (4.5.16); that is,

$$\bar{R}^T B\bar{R} = \begin{bmatrix} B_1 & & & 0 \\ & B_2 & & \\ & & \ddots & \\ 0 & & & B_k \end{bmatrix} = R^*AR\Lambda$$

$$= \begin{bmatrix} A_1\Lambda_1 & & & 0 \\ & A_2\Lambda_2 & & \\ & & \ddots & \\ 0 & & & A_k\Lambda_k \end{bmatrix}$$

with all $B_i, A_i, \Lambda_i \in M_{n_i}$ and $\Lambda_i = \sigma_i D_i^2$ with $\sigma_i \ge 0$,

$$D_j = \operatorname{diag}(e^{i\theta_{1j}}, e^{i\theta_{2j}}, \ldots, e^{i\theta_{n_j j}})$$

and all $\theta_{ij} \in \mathbf{R}$. If $k = n$, this is the required reduction. If $k < n$, then some block has $n_i > 1$ and $B_i = A_i\Lambda_i = \sigma_i A_i D_i^2$. Since D_i is diagonal and unitary, $D_i^* = \bar{D}_i = \bar{D}_i^T = D_i^{-1}$ and hence

$$\bar{D}_i^T B_i \bar{D}_i = \sigma_i D_i^* A_i D_i \tag{4.5.17}$$

The left-hand side of this identity is a symmetric matrix $\bar{D}_i^T B_i \bar{D}_i$, and the right-hand side is a Hermitian matrix $\sigma_i D_i^* A_i D_i$ with σ_i real. If $\sigma_i \neq 0$, we conclude that $D_i^*A_iD_i$ is both Hermitian and symmetric. The only way a Hermitian matrix can be symmetric is if it is real, so $D_i^*A_iD_i$ is a real symmetric matrix if $\sigma_i \neq 0$. If $\sigma_i = 0$ (which can happen for at most one value of i), then $D_i^*A_iD_i$ is Hermitian, but not necessarily real. By the spectral theorem, for each $i = 1, \ldots, k$ there is a unitary matrix $U_i \in M_{n_i}$ and a real diagonal matrix M_i such that $D_i^*A_iD_i = U_iM_iU_i^*$. If $\sigma_i \neq 0$, then U_i may be chosen to be a real orthogonal matrix, in which case $U_i^T = U_i^*$ and

$$\bar{D}_i^T B_i \bar{D}_i = \sigma_i D_i^* A_i D_i = U_i(\sigma_i M_i)U_i^T$$

If $\sigma_i = 0$, then it may not be true that $U_i^* = U_i^T$, but nevertheless the displayed equation is still correct because both sides vanish. Thus, for all $i = 1, 2, \ldots, k$ we have

$$A_i = (D_iU_i)M_i(D_iU_i)^* \quad \text{and} \quad B_i = (D_iU_i)(\sigma_iM_i)(D_iU_i)^T$$

If we set

$$U = \begin{bmatrix} D_1U_1 & & & 0 \\ & D_2U_2 & & \\ & & \ddots & \\ 0 & & & D_kU_k \end{bmatrix}$$

$$M = \begin{bmatrix} M_1 & & & 0 \\ & M_2 & & \\ & & \ddots & \\ 0 & & & M_k \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_1 I & & & 0 \\ & \sigma_2 I & & \\ & & \ddots & \\ 0 & & & \sigma_k I \end{bmatrix}$$

then $B=[(\bar{R}^{-1})^TU]\Sigma M[U^T\bar{R}^{-1}]$ and $A=[(R^{-1})^*U]M[U^*R]$, as asserted. If we assume III(a)(1), then R is unitary and hence $(R^{-1})^*U=RU$ and $(\bar{R}^{-1})^TU=RU$ is unitary, so the sufficiency of III(a)(1) is proved.

This completes the proof of III when A is nonsingular. If B is nonsingular, the hypothesis III(b)(1) says that there is a nonsingular $R\in M_n$ such that $R^{-1}B^{-1}A\bar{R}=\Lambda$ is diagonal, so $A\bar{R}=BR\Lambda$ and $\bar{R}^*A\bar{R}=R^TBR\Lambda$. From this point, the argument is formally the same as in the case in which A is nonsingular. One merely interchanges A and B in the proof and uses the Takagi factorization (4.4.4) to diagonalize $D_i^TB_iD_i$ instead of the spectral theorem. □

In cases I and II of Theorem (4.5.15) (Table 4.5.15T) there is a familiar condition on $A^{-1}B$ that is equivalent to simultaneous diagonalizability by the respective congruence, namely that $A^{-1}B$ is diagonalizable (perhaps with real eigenvalues), that is, $A^{-1}B$ is of the form $R\Lambda R^{-1}$ with Λ diagonal (perhaps with Λ real). This condition can, in principle, be tested by examining the minimal polynomial of $A^{-1}B$ to see if it has distinct linear (perhaps real) factors. In case III, however, the condition is not so familiar, namely that $A^{-1}B$ is of the form $R\Lambda\bar{R}^{-1}$ with Λ diagonal. This requirement says that $A^{-1}B$ is diagonalizable by *consimilarity,* rather than by ordinary similarity. See Section (4.6) for a discussion of consimilarity. Theorem (4.6.11) shows that condition (4.5.15.III(b)(1)) is equivalent to the condition that $C\bar{C}$ has all real nonnegative eigenvalues and is diagonalizable and rank $C=$ rank $C\bar{C}$.

It was convenient to make a nonsingularity assumption in Theorem (4.5.15), but this assumption can be eliminated in the cases I(a), II(a), and III(a) of unitary congruence. In case I(a), this approach gives an alternative proof of the classical result (4.1.6) on simultaneous unitary diagonalization of commuting Hermitian matrices.

4.5.18 **Corollary.** Let $A, B\in M_n$.

(a) If A and B are both Hermitian, there exists a unitary $U\in M_n$ such that UAU^* and UBU^* are both diagonal if and only if AB is Hermitian; that is, $AB=BA$.

(b) If A and B are both symmetric, there exists a unitary $U\in M_n$ such that UAU^T and UBU^T are both diagonal if and only if $A\bar{B}$ is normal; that is, $A\bar{B}B\bar{A}=B\bar{A}A\bar{B}$.

(c) If A is Hermitian and B is symmetric, there exists a unitary $U\in M_n$ such that UAU^* and UBU^T are both diagonal if and only if AB is symmetric; that is, $AB=B\bar{A}$.

Proof: (a) If $UAU^*=\Lambda$ and $UBU^*=M$ are both diagonal, then $A=U^*\Lambda U$, $B=U^*MU$ and $AB=U^*\Lambda UU^*MU=U^*\Lambda MU=U^*M\Lambda U=U^*MUU^*\Lambda U=BA$. Conversely, if $AB=BA$, then $A_\epsilon\equiv A+\epsilon I$ is nonsingular and Hermitian for some $\epsilon>0$ and $A_\epsilon B=(A+\epsilon I)B=AB+\epsilon B=BA+\epsilon B=B(A+\epsilon I)=BA_\epsilon$. Thus, B commutes with A_ϵ and A_ϵ^{-1} and hence $A_\epsilon^{-1}B$ is Hermitian. By I(a)(3) of (4.5.15) (Table 4.5.15T) there exists a unitary U_ϵ such that $U_\epsilon A_\epsilon U_\epsilon^*=U_\epsilon AU_\epsilon^*+\epsilon I=\Lambda_\epsilon$ and $U_\epsilon BU_\epsilon^*=M_\epsilon$ are both diagonal, and hence $U_\epsilon AU_\epsilon^*=\Lambda_\epsilon-\epsilon I$ and $U_\epsilon BU_\epsilon^*=M_\epsilon$ are both diagonal.

(b) If $UAU^T=\Lambda$ and $UBU^T=M$ are both diagonal, then $A=U^*\Lambda\bar{U}$, $B=U^*M\bar{U}$, and $A\bar{B}=U^*\Lambda\bar{U}U^T\bar{M}U=U^*(\Lambda\bar{M})U$ is unitarily diagonalizable and hence normal. For the converse, suppose $A\bar{B}$ is normal and assume that A is nonsingular. Then $A\bar{B}=(A^{-1})^{-1}\bar{B}$ is normal, and II(a)(3) of (4.5.15) says that the two symmetric matrices A^{-1} and $\bar{B}$ are simultaneously unitarily diagonalizable. Thus, there is a unitary $U\in M_n$ and diagonal $\Lambda, M\in M_n$ such that $A^{-1}=U\Lambda U^T$ and $\bar{B}=UMU^T$. Then $A=\bar{U}\Lambda^{-1}\bar{U}^T$ and $B=\bar{U}\bar{M}\bar{U}^T$, which is a simultaneous diagonalization of A and B of the required form. If A is singular, then by (4.4.4) there is a unitary $U\in M_n$ such that UAU^T is diagonal, and we can permute the columns of U if necessary so that

$$UAU^T=\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}, \qquad \Sigma\in M_k, \qquad 1\le k<n$$

and Σ is symmetric (in fact, diagonal) and nonsingular. If we write UBU^T in corresponding block form

$$UBU^T=\begin{bmatrix} B_{11} & B_{12} \\ B_{12}^T & B_{22} \end{bmatrix}, \qquad B_{11}\in M_k, \qquad B_{22}\in M_{n-k}$$

then the blocks B_{11} and B_{22} are symmetric and we have

$$(UAU^T)(\overline{UBU^T})=UA\bar{B}U^*=\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \bar{B}_{11} & \bar{B}_{12} \\ B_{12}^* & \bar{B}_{22} \end{bmatrix}=\begin{bmatrix} \Sigma\bar{B}_{11} & \Sigma\bar{B}_{12} \\ 0 & 0 \end{bmatrix}$$

But $UA\bar{B}U^*$ is normal, too, and hence $\Sigma\bar{B}_{12}=0$ (see Problem 20 at the end of this section) and $B_{12}=0$ since Σ is nonsingular. This shows that

$$UAU^T=\begin{bmatrix}\Sigma & 0\\ 0 & 0\end{bmatrix},\qquad UBU^T=\begin{bmatrix}B_{11} & 0\\ 0 & B_{22}\end{bmatrix},$$

and

$$(UAU^T)(\overline{UBU^T})=\begin{bmatrix}\Sigma\bar{B}_{11} & 0\\ 0 & 0\end{bmatrix}$$

By the previous argument for the nonsingular case, we know that there is a unitary $V_1\in M_k$ and diagonal $\Lambda_1,\Lambda_2\in M_k$ such that $\Sigma=V_1\Lambda_1V_1^T$ and $B_{11}=V_1\Lambda_2V_1^T$. Since B_{22} is symmetric, we also know there is a unitary $V_2\in M_{n-k}$ and a diagonal $\Lambda_3\in M_{n-k}$ such that $B_{22}=V_2\Lambda_3V_2^T$. If we let $\Lambda=\Lambda_1\oplus 0\in M_n$, $M=\Lambda_2\oplus\Lambda_3$, and $V=V_1\oplus V_2$, we have $UAU^T=V\Lambda V^T$, $UBU^T=VMV^T$. Thus, $A=(U^*V)\Lambda(U^*V)^T$ and $B=(U^*V)M(U^*V)^T$ is a simultaneous diagonalization of the required form.

(c) If $UAU^*=\Lambda$ and $UBU^T=M$ are both diagonal, then Λ is necessarily real. We have $A=U^*\Lambda U$, $B=U^*M\bar{U}$, and

$$\begin{aligned}AB&=U^*\Lambda UU^*M\bar{U}=U^*\Lambda M\bar{U}=U^*M\Lambda\bar{U}\\&=U^*M\bar{U}U^T\Lambda\bar{U}=(U^*M\bar{U})(\overline{U^*\Lambda U})=B\bar{A}\end{aligned}$$

Conversely, if $AB=B\bar{A}$, then $A_\epsilon\equiv A+\epsilon I$ is nonsingular and Hermitian for some $\epsilon>0$ and $A_\epsilon B=AB+\epsilon B=B\bar{A}+\epsilon B=B\bar{A}_\epsilon$. Thus, condition III(a)(4) of (4.5.15) is satisfied and there exists a unitary $U_\epsilon\in M_n$ such that $U_\epsilon A_\epsilon U_\epsilon^*=U_\epsilon AU_\epsilon^*+\epsilon I=\Lambda_\epsilon$ and $U_\epsilon BU_\epsilon^T=M_\epsilon$ are both diagonal, and hence $U_\epsilon AU_\epsilon^*=\Lambda_\epsilon-\epsilon I$ and $U_\epsilon BU_\epsilon^T=M_\epsilon$ are both diagonal. □

The problem of simultaneously diagonalizing two singular Hermitian matrices by *congruence (not necessarily unitarily) is considered in Problem 8.

We have seen that, under *congruence, a single Hermitian matrix may always be brought to a remarkably simple form (diagonal with ±1 or 0's on the diagonal), and, under certain conditions, a pair of Hermitian matrices may be brought simultaneously into diagonal form by *congruence. A natural question to raise, then, is: Into what canonical form may a pair of general Hermitian matrices $A,B\in M_n$ be brought under *simultaneous* *congruence? That is, what canonical forms may the pair

$$C^*AC \quad\text{and}\quad C^*BC$$

assume with a single congruence by C? Although this question has been treated for general Hermitian pairs (with possibly both singular), the general result is significantly complicated, both to present and to prove. We state without proof here the canonical pair theorem for Hermitian

matrices, which covers the case in which at least one of the matrices is nonsingular. We have already treated the special case in which simultaneous diagonalization by *congruence is possible.

4.5.19 **Theorem.** Suppose that $A,B\in M_n$ are Hermitian, and that A is nonsingular. There is a positive integer k and a nonsingular $C\in M_n$ such that

$$C^*AC=\begin{bmatrix}A_1 & & & 0\\ & A_2 & & \\ & & \ddots & \\ 0 & & & A_k\end{bmatrix},\qquad C^*BC=\begin{bmatrix}B_1 & & & 0\\ & B_2 & & \\ & & \ddots & \\ 0 & & & B_k\end{bmatrix}$$

where each pair $A_i,B_i\in M_{n_i}$, $i=1,2,\ldots,k$ is one of the two possible forms:

$$B_i=\epsilon\begin{bmatrix}0 & & & \alpha\\ & & \cdot^{\cdot^{\cdot}} & 1\\ & \cdot^{\cdot^{\cdot}} & \cdot^{\cdot^{\cdot}} & \\ \alpha & 1 & & 0\end{bmatrix},\qquad A_i=\epsilon\begin{bmatrix}0 & & 1\\ & \cdot^{\cdot^{\cdot}} & \\ 1 & & 0\end{bmatrix}\tag{4.5.20}$$

with α real, or

$$B_i=\left[\begin{array}{c|c}0 & \begin{matrix}0 & & \alpha\\ & \cdot^{\cdot^{\cdot}} & 1\\ \alpha\;\;1 & & 0\end{matrix}\\ \hline \begin{matrix}0 & & \bar{\alpha}\\ & \cdot^{\cdot^{\cdot}} & 1\\ \bar{\alpha}\;\;1 & & 0\end{matrix} & 0\end{array}\right],\qquad A_i=\begin{bmatrix}0 & & 1\\ & \cdot^{\cdot^{\cdot}} & \\ 1 & & 0\end{bmatrix}\tag{4.5.21}$$

with α complex. In (4.5.20), ϵ is either ±1, and in (4.5.21), n_i is even and the two nonzero blocks are both in $M_{(1/2)n_i}$.

Notes

1. In the case that α is real, $n_i=1$ is possible, and the two blocks are then of the form $\pm\alpha$, ±1. Several 1-by-1 blocks corresponding to the same value of α (and with the same value $\epsilon=1$, say) would produce a block of the form αI in C^*BC and I in C^*AC.

2. In the case that α is complex, $n_i = 2$ is possible, and the two blocks then are of the form

$$B_i = \begin{bmatrix} 0 & \alpha \\ \bar{\alpha} & 0 \end{bmatrix}, \qquad A_i = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

3. In the circumstance of the theorem, the simultaneous block structure corresponds exactly to the Jordan canonical form of $A^{-1}B$. That is, the basic Jordan blocks of $A^{-1}B$ are precisely the $A_i^{-1}B_i$. Note that $(C^*AC)^{-1}(C^*BC) = C^{-1}(A^{-1}B)C$, so that C is also a similarity matrix that takes $A^{-1}B$ to Jordan canonical form. Thus the form guaranteed by the theorem can be found from the Jordan canonical form of $A^{-1}B$ (determination of the inertial factors, the ϵ terms, is auxiliary).

4.5.22 **Remark.** Just as the canonical pair form (4.5.19) for two Hermitian matrices A, B under *congruence is analogous to the Jordan canonical form of $A^{-1}B$, there is a canonical pair form for two *real* symmetric matrices A, B under *real* ^{T}congruence that is analogous to the *real* Jordan canonical form of $A^{-1}B$. In it, the blocks B_i of type (4.5.21) are replaced by the natural analog of blocks of the type (3.4.4), while other possible block types are the same.

Problems

1. Let $A, B \in M_n$ and suppose B is nonsingular. Show that there exists $C \in M_n$ such that $A = BC$. Moreover, for any nonsingular $S \in M_n$, we have $SAS^* = (SBS^*)C'$, where C' is similar to C.

2. The only nontrivial part of the proof of Sylvester's law of inertia (4.5.8) is to show that if D_1 and D_2 are congruent n-by-n inertia matrices (4.5.7), then they have exactly the same number of positive diagonal entries. The argument given in the text relies on a corollary of the Courant-Fischer theorem. Provide the details for the following elementary argument: Suppose $D_2 = S^*D_1S$, and suppose D_1 has exactly s positive diagonal entries and at least one negative diagonal entry. Suppose that the first s diagonal entries of D_1 and the first t diagonal entries of D_2 are positive with $1 \le s, t < n$. If $s < t$, show that there is a nonzero vector $x = [x_i] \in \mathbf{C}^n$ such that $x_{t+1} = x_{t+2} = \cdots = x_n = 0$ and $(Sx)_1 = (Sx)_2 = \cdots = (Sx)_s = 0$. Then show that $x^*D_2x > 0$ and $(Sx)^*D_1(Sx) < 0$ to obtain a contradiction.

3. Let $A, B \in M_n$ be Hermitian. Show that the following four conditions are equivalent: (a) A and B are simultaneously diagonalizable by *congruence. (b) For some real nonzero scalars a, b, $aA + bB$ and B are simultaneously diagonalizable by *congruence. (c) A and B are simultaneously *congruent to a pair of commuting matrices. (d) $A + iB$ is *congruent to a normal matrix.

4. Use the argument in the proof of (4.5.15) and the commuting family Theorems (1.3.19) and (4.1.6) to prove the following generalization of I(b) of Theorem (4.5.15). Let $A_1, A_2, \ldots, A_k \in M_n$ be given Hermitian matrices with A_1 nonsingular. There exists a nonsingular matrix $T \in M_n$ such that T^*A_iT is diagonal for all $i = 1, 2, \ldots, k$ if and only if (a) $A_1^{-1}A_i$ is similar to a real diagonal matrix for all $i = 2, \ldots, k$, and (b) $\{A_1^{-1}A_i : i = 2, \ldots, n\}$ is a commuting family of matrices. *Hint:* Let $C_i = A_1^{-1}A_i$ and let SC_iS^{-1} be real diagonal for all $i = 2, \ldots, k$. Let $B_i = (S^*)^{-1}A_iS^{-1}$ and show that $\{B_i\}$ is a commuting family of Hermitian matrices. There is a unitary matrix U such that UB_iU^* is diagonal for all $i = 2, \ldots, k$ and $T = US$ is the required congruence matrix. What is the corresponding generalization of II(b) of (4.5.15)?

5. A differential operator L given by (4.0.4) with a real symmetric coefficient matrix $A(x) = [a_{ij}(x)]$ is said to be *elliptic* at a point $x \in D \subset \mathbf{R}^n$ if the coefficient matrix $A(x)$ is nonsingular and all its eigenvalues have the same sign. L is said to be ***hyperbolic*** at x if $A(x)$ is nonsingular, $n-1$ of the eigenvalues have the same sign, and one eigenvalue has the opposite sign. Explain why a differential operator that is elliptic (or hyperbolic) at a point with respect to one coordinate system is elliptic (or hyperbolic) at that point with respect to every other coordinate system. Laplace's equation

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = 0$$

gives an example of an elliptic differential operator, and the wave equation

$$\Box^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} - \frac{\partial^2 f}{\partial t^2} = 0$$

is an example of a hyperbolic one. Both are presented in Cartesian coordinates. Both look very different in spherical polar, cylindrical, or other coordinates.

6. Let $X = [X_1, \ldots, X_n]^T$ and $Y = [Y_1, \ldots, Y_n]^T$ be two vectors of real random variables with finite second moments. It is a fact (see Chapter 7) that the covariance matrices of X and Y each have only nonnegative eigenvalues. Suppose that at least one of the covariance matrices is nonsingular. Show that there exists a real nonsingular matrix $S \in M_n$ such that the covariance matrices of SX and SY are both diagonal. In statistical

terms, this says that a single nonsingular linear transformation S can be found so that the components of SX and SY are each uncorrelated.

7. Use Problem 4 to give conditions on three or more random vectors that are sufficient to guarantee the existence of a single nonsingular linear transformation with the property that the components of the transformed random vectors are uncorrelated.

8. Case I(b) of Theorem (4.5.15) considers the problem of simultaneous diagonalization of two Hermitian matrices by *congruence in the case that at least one of the matrices is nonsingular. Corollary (4.5.18a) considers the problem of simultaneous diagonalization by *unitary* *congruence when both matrices may be singular. If the two matrices are both singular, the problem of simultaneously diagonalizing them by (not necessarily unitary) *congruence can be reduced eventually to (4.5.15), but one must look at the behavior of the two matrices on the orthogonal complement of the intersection of their null spaces. Let $A, B \in M_n$ be Hermitian and assume that both are singular. Let $N(A)$ and $N(B)$ denote the null spaces of A and B, respectively. (a) Consider $\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ and $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ to show that there exist pairs of singular Hermitian matrices that are simultaneously diagonalizable by *congruence. (b) Suppose $N(A) \cap N(B) = \{0\}$. Show that if A and B are simultaneously diagonalizable by *congruence, then there exists a real number a such that $aA+B$ is nonsingular. *Hint:* If $C \in M_n$ is nonsingular, $C^*AC = \Lambda_1$, and $C^*BC = \Lambda_2$ with Λ_1 and Λ_2 diagonal, show that the zero main diagonal entries of Λ_1 and Λ_2 do not fall in the same positions. Can you select a so that all the main diagonal entries of $a\Lambda_1 + \Lambda_2$ are nonzero? (c) Use (b) to show that

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

cannot be diagonalized simultaneously by *congruence. (d) If

$$N(A) \cap N(B) = \{0\}$$

if $a \in \mathbf{R}$ is nonzero, and if $aA+B$ is nonsingular, use Problem 3(b) to show that A and B are simultaneously diagonalizable by *congruence if and only if $(aA+B)^{-1}B$ is diagonalizable and has only real eigenvalues. (e) If $\dim N(A) \cap N(B) = k \geq 1$, let $\{u_1, u_2, \ldots, u_n\}$ be an orthonormal basis of $\mathbf{R}^n$ for which $\{u_1, u_2, \ldots, u_k\}$ is an orthonormal basis of $N(A) \cap N(B)$. If $U = [u_1\, u_2 \cdots u_n] \in M_n$, show that

$$U^*AU = \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & A' \end{array}\right] \quad \text{and} \quad U^*BU = \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & B' \end{array}\right]$$

where $A', B' \in M_{n-k}$, $N(A') \cap N(B') = \{0\}$, and the upper left-hand corner zero blocks are k-by-k. Show that A and B are simultaneously diagonalizable by *congruence if and only if A' and B' are simultaneously diagonalizable by *congruence. Although A' and B' may both be singular, the intersection of their null spaces is trivial. (f) Assemble the information from (a)–(e) to state and prove a general theorem on simultaneous diagonalization of two Hermitian matrices by *congruence.

9. If $A, B \in M_n$ and B is nonsingular, show that A commutes with B if and only if A commutes with B^{-1}.

10. Show that $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ and $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ can be reduced simultaneously to diagonal form by a unitary ^{T}congruence but cannot be reduced simultaneously to diagonal form by *congruence. Use the construction employed in the proof of case II(b) of (4.5.15) to carry out the reduction and find an explicit unitary ^{T}congruence matrix which works.

11. Show that $\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ cannot be reduced simultaneously to diagonal form by either *congruence or ^{T}congruence.

12. Let $A, B \in M_n$ with A nonsingular. Show that each of the following conditions is necessary and sufficient for A and B to be simultaneously diagonalizable by congruence in the sense of, and under the assumptions of, each of the indicated cases of Theorem (4.5.15) (Table 4.5.15T).

Case	Necessary and sufficient condition
I(a)	There exists a Hermitian $F \in M_n$ such that $B = AF$.
I(b)	There exists a diagonalizable $F \in M_n$ with real eigenvalues such that $B = AF$.
II(a)	There exists a normal $F \in M_n$ such that $B = AF$.
II(b)	There exists a diagonalizable $F \in M_n$ such that $B = AF$.
III(a)	There exists a symmetric $F \in M_n$ such that $B = AF$.
III(b)	There exists a condiagonalizable matrix $F \in M_n$ such that $B = AF$ [see (4.6.2)].

13. Let $A, B \in M_n$ be symmetric (possibly both singular) and suppose that there is a unitary $U \in M_n$ such that $UAU^T = \Lambda$ and $UBU^T = M$ are both diagonal. Show that there exists a unitary matrix V such that $B\bar{A} =$

$AV\bar{B}$. *Hint:* If $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, show that there exists a unitary diagonal matrix D such that $\bar{\Lambda} = D\Lambda = \Lambda D$. Then show that

$$B\bar{A} = U^*M\bar{\Lambda}U = U^*\Lambda D_1 D_2 \bar{M}U = A(U^T D_1 D_2 \bar{U})\bar{B}$$

where D_1 and D_2 are unitary diagonal matrices.

14. Use the necessary condition in Problem 13 to show that the two symmetric matrices in Problem 8(c) cannot be diagonalized simultaneously by unitary ^{T}congruence. *Hint:* Compute the first column of $B\bar{A}$ and of AUB. Use (4.5.18b) to show the same thing more easily.

15. If $A, B \in M_n$ are symmetric, show that the necessary condition for simultaneous diagonalization by unitary ^{T}congruence in Problem 13 is also a sufficient condition provided that *both* A and B are nonsingular. *Hint:* If $B\bar{A} = AU\bar{B}$ with both A and B nonsingular, then $A^{-1}B\bar{A}\bar{B}^{-1} = U$ and $I = UU^*$. This implies that $\bar{A}\bar{B}^{-1}B^{-1}A = B^{-1}A\bar{A}\bar{B}^{-1}$. Taking the inverse of both sides implies that $A^{-1}B$ is normal.

16. Let $A, B \in M_n$ be symmetric (possibly both singular) and suppose that there is a unitary $U \in M_n$ such that $UAU^T = \Lambda$ and $UBU^T = M$ are both diagonal. Show that $A\bar{A}$ commutes with $B\bar{B}$. Show that this necessary condition for simultaneous diagonalization by unitary ^{T}congruence is *not* sufficient by considering the two matrices in Problem 8(c). Use Corollary (4.4.5) to show that this necessary condition *is* sufficient provided that both $A\bar{A}$ and $B\bar{B}$ have n distinct eigenvalues.

17. Let $A, B \in M_n$ with A Hermitian and B symmetric and suppose there is a unitary $U \in M_n$ such that $UAU^* = \Lambda$ and $UBU^T = M$ are both diagonal. Show that A commutes with $B\bar{B}$. Show that this necessary condition for simultaneous diagonalization by (mixed * and ^{T}) congruence is *not* sufficient by considering the two matrices in Problem 11. Use Corollary (4.4.5) to show that this necessary condition *is* sufficient provided that all the eigenvalues of $B\bar{B}$ are distinct.

18. Let $A, B \in M_n$ with A and B symmetric and A nonsingular. Show that if the generalized characteristic polynomial $p_{A,B}(t) \equiv \det(tA - B)$ has n distinct zeroes, then A and B are simultaneously diagonalizable by ^{T}congruence. *Hint:* What are the eigenvalues of $A^{-1}B$?

19. Provide the details for the following alternative proof of Sylvester's law of inertia (4.5.8). If $A \in M_n$ is Hermitian and nonsingular and if $S \in M_n$ is nonsingular, let $S = QR$ be a factorization in which $Q \in M_n$ is unitary and $R \in M_n$ is upper triangular (2.6.1) with positive main diagonal. Show that $S(t) = tQ + (1-t)QR$ is nonsingular if $0 \le t \le 1$ and let $A(t) \equiv S(t)AS(t)^*$. What is $A(0)$? $A(1)$? Since $A(t)$ is nonsingular and changes continuously as t goes from zero to 1, argue that $A(0)$ has the same number of positive (negative) eigenvalues as $A(1)$. Treat the general case by considering $A \pm \epsilon I$ for small $\epsilon > 0$.

20. If $A = \begin{bmatrix} B & C \\ 0 & 0 \end{bmatrix} \in M_n$ with $B \in M_k$, $1 \le k < n$, show that A is normal if and only if B is normal and $C = 0$. *Hint:* Compute AA^* and A^*A. If $C^*C = 0$, then $(Cx)^*(Cx) = 0$ for all $x \in \mathbf{C}^{n-k}$ and hence $Cx = 0$ for all $x \in \mathbf{C}^{n-k}$.

21. Show that the method of proof used in (b) of (4.5.18) can also be used to prove parts (a) and (c).

22. Let $\mathfrak{F} = \{A_1, \ldots, A_k\} \subset M_n$ be a given family of complex symmetric matrices and let $\mathcal{G} = \{A_i\bar{A}_j : i, j = 1, 2, \ldots, k\}$. If there exists a unitary $U \in M_n$ such that UA_iU^T is diagonal for all $i = 1, \ldots, k$, show that $\mathcal{G}$ is a commuting family. What does this reduce to when $k = 2$, and what is the connection with (4.5.18b)? In fact, commutativity of $\mathcal{G}$ is also sufficient to ensure the simultaneous diagonalizability of $\mathfrak{F}$ by unitary ^{T}congruence; see the paper of Hong and Horn listed in the references at the end of this section.

23. Let $\mathfrak{F} = \{A_1, \ldots, A_k\} \subset M_n$ be a given family of complex symmetric matrices, let $\mathcal{H} = \{B_1, \ldots, B_m\} \subset M_n$ be a given family of Hermitian matrices, and let $\mathcal{G} = \{A_i\bar{A}_j : i, j = 1, \ldots, k\}$. If there is a unitary $U \in M_n$ such that every UA_iU^T and every UB_jU^* is diagonal, show that each of $\mathcal{G}$ and $\mathcal{H}$ is a commuting family and B_jA_i is symmetric for all $i = 1, \ldots, k$ and all $j = 1, \ldots, m$. What does this reduce to when $k = m = 1$, and what is the connection with (4.5.18c)? In fact, these conditions are also sufficient to ensure the simultaneous diagonalizability of $\mathfrak{F}$ and $\mathcal{H}$ by the respective congruences; see the paper of Hong and Horn listed in the references at the end of this section.

Further Readings. Ostrowski's proof of (4.5.9) and related results are in "A Quantitative Formulation of Sylvester's Law of Inertia," *Proc. Nat. Acad. Sci.* 45 (1959), 740–744. Another version of Theorem (4.5.25) is stated in [GLR 82]; a careful proof including the case in which both matrices are singular is contained in unpublished notes by R. C. Thompson. For results about simultaneous diagonalization of more than two matrices, see Y. P. Hong and R. A. Horn, "On Simultaneous Reduction of Families of Matrices to Triangular or Diagonal Form by Unitary Congruence," *Linear and Multilinear Algebra* 17 (1985), 271–288.

4.6 Consimilarity and condiagonalization

The motivation for the topic of this section comes from three results in the preceding two sections. Theorem (4.4.3) characterizes all matrices of the form $U\Delta U^T$, where Δ is upper triangular and U is unitary; for our present purposes we prefer to write this factorization as $U\Delta U^T = U\Delta\bar{U}^{-1}$. Corollary (4.4.4) characterizes all matrices of the form $U\Sigma U^T = U\Sigma\bar{U}^{-1}$, where Σ is diagonal, and case III of Theorem (4.5.15) requires information about when a given square complex matrix A can be reduced to diagonal form by the transformation $A \to SA\bar{S}^{-1}$ for some nonsingular S.

4.6.1 **Definition.** Two matrices $A, B \in M_n$ are said to be *consimilar* if there exists a nonsingular $S \in M_n$ such that $A = SB\bar{S}^{-1}$. If the matrix S can be taken to be unitary, A and B are said to be *unitarily consimilar.*

If $A = SB\bar{S}^{-1}$ and $S = U$ is unitary, then $A = SB\bar{S}^{-1} = UBU^T$; if $S = Q$ is complex orthogonal, then $A = SB\bar{S}^{-1} = QBQ^*$; if $S = R$ is a real nonsingular matrix, then $A = SB\bar{S}^{-1} = RBR^{-1}$. Thus, special cases of consimilarity include ^{T}congruence, *congruence, and ordinary similarity.

Like ordinary similarity, consimilarity is an equivalence relation on M_n, and we may ask which equivalence classes contain triangular or diagonal representatives.

4.6.2 **Definition.** A matrix $A \in M_n$ is said to be *contriangularizable* if there exists a nonsingular $S \in M_n$ such that $S^{-1}A\bar{S}$ is upper triangular; it is said to be *condiagonalizable* if S can be chosen so that $S^{-1}A\bar{S}$ is diagonal. It is said to be *unitarily contriangularizable* or *unitarily condiagonalizable* if it can be reduced by consimilarity to the required form via a unitary matrix.

If $A \in M_n$ is contriangularizable, and if $S^{-1}A\bar{S} = \Delta$ is upper triangular, then an explicit calculation shows that the main diagonal entries of $\Delta\bar{\Delta} = S^{-1}(A\bar{A})S$ are nonnegative. Consequently, all the eigenvalues of $A\bar{A}$ are nonnegative. But then Theorem (4.4.3) says that there is a unitary U such that $UAU^T = UA\bar{U}^{-1}$ is upper triangular. Thus, the problem of deciding whether a given matrix can be reduced to upper triangular form by consimilarity has already been solved.

4.6.3 **Theorem.** Let $A \in M_n$ be given. The following statements are equivalent:

(a) A is contriangularizable;
(b) A is unitarily contriangularizable; and
(c) All the eigenvalues of $A\bar{A}$ are real and nonnegative.

If $A \in M_n$ is unitarily condiagonalizable, then $A = U\Lambda\bar{U}^{-1} = U\Lambda U^T$ for some unitary $U \in M_n$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$. Thus, $A^T = (U\Lambda U^T)^T = U\Lambda^T U^T = U\Lambda U^T = A$, and hence A is symmetric. Corollary (4.4.4) says that the converse is true as well, and that the diagonal matrix can always be taken to be nonnegative. Thus, the problem of unitary condiagonalization has also been solved already.

4.6.4 **Theorem.** A matrix $A \in M_n$ is unitarily condiagonalizable if and only if it is symmetric.

The remaining problem concerning contriangularization and condiagonalization is to characterize usefully those matrices that can be condiagonalized by a consimilarity that is not necessarily unitary.

If $A \in M_n$ is condiagonalizable and $S^{-1}A\bar{S} = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$, then $A\bar{S} = S\Lambda$. If $S = [s_1 \ldots s_n]$ with each $s_i \in \mathbf{C}^n$, this identity says that $A\bar{s}_i = \lambda_i s_i$ for $i = 1, \ldots, n$. This equation is similar to, but crucially different from, the usual eigenvector–eigenvalue equation.

4.6.5 **Definition.** Let $A \in M_n$ be given. A nonzero vector $x \in \mathbf{C}^n$ such that $A\bar{x} = \lambda x$ for some $\lambda \in \mathbf{C}$ is said to be a *coneigenvector* of A; the scalar λ is a *coneigenvalue* of A.

The identity $A\bar{S} = S\Lambda$ says that every nonzero column of S is a coneigenvector of A. Since the columns of S are independent if and only if S is nonsingular, we see that a matrix $A \in M_n$ is condiagonalizable if and only if it has n independent coneigenvectors. To this extent, the theory of condiagonalization is entirely analogous to the theory of ordinary diagonalization.

But every matrix has at least one eigenvalue, and it has only finitely many distinct eigenvalues; in this regard, the theory of coneigenvalues is rather different. If $A\bar{x} = \lambda x$, then $e^{-i\theta}A\bar{x} = A(\overline{e^{i\theta}x}) = e^{-i\theta}\lambda x = (e^{-2i\theta}\lambda)(e^{i\theta}x)$ for all $\theta \in \mathbf{R}$. Thus, if λ is a coneigenvalue of A, then so is $e^{i\theta}\lambda$ for all $\theta \in \mathbf{R}$. On the other hand, if $A\bar{x} = \lambda x$, then $A\bar{A}x = A(\overline{A\bar{x}}) = A(\overline{\lambda x}) = \bar{\lambda}A\bar{x} = \bar{\lambda}\lambda x = |\lambda|^2 x$, so a scalar λ is a coneigenvalue of A only if $|\lambda|^2$ is an eigenvalue of $A\bar{A}$. The example $A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$, for which $A\bar{A} = -2I$ has no nonnegative eigenvalues, shows that there are matrices that have no coneigenvalues at all. It is known, however, that if $A \in M_n$ and n is odd, then A must have at least one coneigenvalue, a result analogous

to the fact that every real matrix of odd order has at least one real eigenvalue.

Thus, in contrast to the theory of ordinary eigenvalues, a matrix may have infinitely many distinct coneigenvalues or it may have no coneigenvalues at all. If a matrix has a coneigenvalue, it is sometimes convenient to select from among the coneigenvalues of equal modulus the unique nonnegative one as a representative.

The necessary condition we have just observed for the existence of a coneigenvalue is also sufficient.

4.6.6 **Proposition.** Let $A \in M_n$ and let $\lambda \geq 0$ be given. Then λ is an eigenvalue of $A\bar{A}$ if and only if $+\sqrt{\lambda}$ is a coneigenvalue of A.

Proof: If $\lambda \geq 0$, $\sqrt{\lambda} \geq 0$, and $A\bar{x} = \sqrt{\lambda}x$ for some $x \neq 0$, then $A\bar{A}x = A(\overline{A\bar{x}}) = A(\overline{\sqrt{\lambda}x}) = \sqrt{\lambda}A\bar{x} = \sqrt{\lambda}\sqrt{\lambda}x = \lambda x$.

Conversely, if $A\bar{A}x = \lambda x$ for some $x \neq 0$, there are two possibilities:

(a) $A\bar{x}$ and x are dependent; or
(b) $A\bar{x}$ and x are independent.

In the former case, there is some $\mu \in \mathbf{C}$ such that $A\bar{x} = \mu x$, which says that μ is a coneigenvalue of A. But then $\lambda x = A\bar{A}x = A(\overline{A\bar{x}}) = A(\overline{\mu x}) = \bar{\mu}A\bar{x} = \bar{\mu}\mu x = |\mu|^2 x$, so $|\mu| = +\sqrt{\lambda}$. Since $e^{-2i\theta}\mu$ is a coneigenvalue associated with the coneigenvector $e^{i\theta}x$ for any $\theta \in \mathbf{R}$, we conclude that $+\sqrt{\lambda}$ is a coneigenvalue of A. Notice that $A\bar{A}(A\bar{x}) = A(\overline{A\bar{A}x}) = A(\overline{\lambda x}) = \lambda(A\bar{x})$ and $A\bar{A}x = \lambda x$, so if λ is a simple eigenvalue of $A\bar{A}$, (a) must always be the case.

In the latter case (b) (which could occur if λ is a multiple eigenvalue of $A\bar{A}$), the vector $y = A\bar{x} + \sqrt{\lambda}x$ is nonzero and is a coneigenvector corresponding to the coneigenvalue $+\sqrt{\lambda}$ since

$$A\bar{y} = A\bar{A}x + \sqrt{\lambda}A\bar{x} = \lambda x + \sqrt{\lambda}A\bar{x} = \sqrt{\lambda}(A\bar{x} + \sqrt{\lambda}x) = \sqrt{\lambda}y \qquad \square$$

We have seen that to each distinct nonnegative eigenvalue of $A\bar{A}$ there corresponds a coneigenvector of A, a result analogous to the ordinary theory of eigenvectors. The following result extends this analogy a bit further.

4.6.7 **Proposition.** Let $A \in M_n$ be given, and let $x_1, x_2, \ldots, x_k$ be coneigenvectors of A with corresponding coneigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$. If $|\lambda_i| \neq |\lambda_j|$ whenever $1 \leq i, j \leq k$ and $i \neq j$, then $\{x_1, \ldots, x_k\}$ is an independent set.

Proof: Each x_i is an eigenvector of $A\bar{A}$ with associated eigenvalue $|\lambda_i|^2$. The vectors $x_1, \ldots, x_k$ are independent by (1.3.8) because they are eigenvectors of the matrix $A\bar{A}$ and their associated eigenvalues $|\lambda_1|^2, \ldots, |\lambda_k|^2$ are distinct by assumption. $\square$

This result, together with Proposition (4.6.6), gives a lower bound on the number of independent coneigenvectors of a given matrix and yields a sufficient condition for condiagonalizability that is analogous to a familiar sufficient condition for ordinary diagonalizability. We give a more general condition in Theorem (4.6.11).

4.6.8 **Corollary.** Let $A \in M_n$ be given. If $A\bar{A}$ has k distinct nonnegative eigenvalues, then A has at least k independent coneigenvectors. If $k = n$, A is condiagonalizable. If $k = 0$, A has no coneigenvectors at all.

These bounds on the number of independent coneigenvectors are sharp. For $A = J_n(1)$, an elementary Jordan block

$$J_n = \begin{bmatrix} 1 & 1 & & 0 \\ & 1 & \ddots & \\ & & \ddots & 1 \\ 0 & & & 1 \end{bmatrix} \in M_n$$

$A\bar{A} = J_n^2(1)$ has 1 as its only nonnegative eigenvalue. The coneigenvector equation $A\bar{x} = x$ is easily seen to have only real solutions, so every coneigenvector is also an eigenvector, and the subspace of eigenvectors is one-dimensional. Direct sums of elementary Jordan blocks can therefore be used to give examples of matrices $A \in M_n$ such that $A\bar{A}$ has k distinct nonnegative eigenvalues and A has exactly k independent coneigenvectors for any integer k such that $1 \leq k \leq n$.

Our objective is to give a simple condition for a given matrix to be condiagonalizable, and as a first step we prove the following lemma. The motivation for this result is that if a given matrix $A \in M_n$ is consimilar to a scalar matrix, then $A = S(\lambda I)\bar{S}^{-1} = \lambda S\bar{S}^{-1}$ and $A\bar{A} = \lambda S\bar{S}^{-1}\bar{\lambda}\bar{S}S^{-1} = |\lambda|^2 I$. Matrices with this property (that $A\bar{A}$ is a scalar matrix) are the basic building blocks from which condiagonalizable matrices are constructed.

4.6.9 **Lemma.** A matrix $A \in M_n$ has the property that $A\bar{A} = I$ if and only if there exists a nonsingular $S \in M_n$ such that $A = S\bar{S}^{-1}$.

Proof: We have just seen that the stated condition is necessary. To show that it is sufficient, define $S_\theta \equiv e^{i\theta}A + e^{-i\theta}I$ for any $\theta \in \mathbf{R}$ and observe that

$$A\bar{S}_\theta = A(e^{-i\theta}\bar{A}+e^{i\theta}I) = e^{-i\theta}A\bar{A}+e^{i\theta}A = e^{i\theta}A+e^{-i\theta}I = S_\theta \qquad (4.6.10)$$

Since A has only finitely many eigenvalues, there is some $\theta_0 \in \mathbf{R}$ such that $-e^{-2i\theta_0}$ is not an eigenvalue of A. For this value of θ,

$$S_{\theta_0} = e^{i\theta}(A+e^{-2i\theta_0}I)$$

is nonsingular and $A = S_{\theta_0}\bar{S}_{\theta_0}^{-1}$ from (4.6.10). □

We can now state and prove a necessary and sufficient condition for condiagonalizability.

4.6.11 **Theorem.** Let $A \in M_n$. There exists a nonsingular $S \in M_n$ and a diagonal $\Lambda \in M_n$ such that $A = S\Lambda\bar{S}^{-1}$ if and only if $A\bar{A}$ is a diagonalizable matrix with real nonnegative eigenvalues and $\operatorname{rank} A = \operatorname{rank} A\bar{A}$.

Proof: The stated conditions are clearly necessary since

$$A\bar{A} = S\Lambda\bar{S}^{-1}\bar{S}\bar{\Lambda}S^{-1} = S|\Lambda|^2 S^{-1}$$

and the rank of both $A\bar{A}$ and A is the number of nonzero diagonal entries in Λ. Conversely, if $A\bar{A}$ is diagonalizable and has nonnegative eigenvalues, there is a nonsingular $S \in M_n$ and a nonnegative diagonal $\Lambda \in M_n$ such that $A\bar{A} = S\Lambda S^{-1}$. There is no loss of generality to assume that like diagonal entries in Λ are grouped together and that $\Lambda = \lambda_1 I_{n_1} \oplus \lambda_2 I_{n_2} \oplus \cdots \oplus \lambda_k I_{n_k}$, where $I_{n_i} \in M_{n_i}$ and $\lambda_1 > \lambda_2 > \lambda_3 > \cdots > \lambda_k \geq 0$. We then have

$$S^{-1}A\bar{A}S = S^{-1}A\bar{S}\bar{S}^{-1}\bar{A}S = (S^{-1}A\bar{S})(\overline{S^{-1}A\bar{S}}) = \Lambda$$

If we set $B = S^{-1}A\bar{S}$, then (since consimilarity is an equivalence relation) it will suffice to show that B is condiagonalizable if $B\bar{B} = \Lambda$. Since Λ is real, $\Lambda = \bar{\Lambda} = (\overline{B\bar{B}}) = \bar{B}B = B\bar{B}$, so B and $\bar{B}$ commute. Thus, $B\Lambda = B(B\bar{B}) = BB\bar{B} = (B\bar{B})B = \Lambda B$, so B and Λ also commute. If we write B in block form as

$$B = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1k} \\ \vdots & B_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ B_{k1} & \cdots & \cdots & B_{kk} \end{bmatrix}$$

with block sizes conformal to those of

$$\Lambda = \begin{bmatrix} \lambda_1 I_{n_1} & & 0 \\ & \ddots & \\ 0 & & \lambda_k I_{n_k} \end{bmatrix}, \qquad I_{n_i} \in M_{n_i}, \qquad i = 1, 2, \ldots, k$$

then the equation $B\Lambda = \Lambda B$ says that $\lambda_i B_{ij} = \lambda_j B_{ij}$ for all $i = 1, 2, \ldots, k$. Since $\lambda_i \neq \lambda_j$ if $i \neq j$, we conclude that $B_{ij} = 0$ if $i \neq j$, and hence B is block diagonal

$$B = \begin{bmatrix} B_{11} & & 0 \\ & \ddots & \\ 0 & & B_{kk} \end{bmatrix}$$

with diagonal blocks the same size as those of Λ. The equation $B\bar{B} = \Lambda$ means that each $B_{ii}\bar{B}_{ii} = \lambda_i I$ for $i = 1, 2, \ldots, k$. Notice that B_{ii} must be nonsingular if $\lambda_i > 0$. If $\lambda_i > 0$, we can write this equation as

$$\left[\frac{1}{\sqrt{\lambda_i}} B_{ii}\right]\left[\overline{\frac{1}{\sqrt{\lambda_i}} B_{ii}}\right] = I_{n_i}$$

and we can use Lemma (4.6.9) to conclude that there is a nonsingular $S_i \in M_{n_i}$ such that $B_{ii} = S_i(\sqrt{\lambda_i} I_{n_i})\bar{S}_i^{-1}$. If $\lambda_k = 0$, then

$$\operatorname{rank} B_{11} + \operatorname{rank} B_{22} + \cdots + \operatorname{rank} B_{kk}$$
$$= \operatorname{rank} B = \operatorname{rank} A = \operatorname{rank} A\bar{A} = \operatorname{rank} \Lambda = n_1 + n_2 + \cdots + n_{k-1}$$

This means that the rank of $B_{k,k}$ is zero, so the last block B_{kk} must actually be a zero block if $\lambda_k = 0$. In this event, we can write $0 = B_{kk} = S_k(\sqrt{\lambda_k} I)\bar{S}_k^{-1}$, where $S_k \in M_{n_k}$ is an arbitrary nonsingular matrix. If we set $S = S_1 \oplus \cdots \oplus S_k$, we have shown in all cases that

$$B = S(\sqrt{\lambda_1} I_{n_1} \oplus \cdots \oplus \sqrt{\lambda_k} I_{n_k})\bar{S}^{-1}$$

and we are done. □

When applied to case III(b) of Theorem (4.5.15), the necessary and sufficient conditions for condiagonalizability have the following consequence. Let $A, B \in M_n$ be given with A Hermitian and B symmetric and at least one of A, B nonsingular. Set $C = A^{-1}B$ or $B^{-1}A$ depending on which is nonsingular. There exists a nonsingular $S \in M_n$ such that both SAS^* and SBS^T are diagonal if and only if $C\bar{C}$ has all nonnegative eigenvalues and is diagonalizable and $\operatorname{rank} C = \operatorname{rank} C\bar{C}$.

The special case in which A is a complex symmetric matrix is handled easily by the theorem, since $A\bar{A} = AA^*$ is Hermitian in this case and hence is diagonalizable. Moreover, $\operatorname{rank} A = \operatorname{rank} AA^*$ for any $A \in M_n$, so the hypotheses of the theorem are satisfied when A is a complex symmetric matrix. The theorem shows that every complex symmetric matrix is condiagonalizable but does not yield directly the fact that the condiagonalization can be accomplished with a unitary transformation in this case. See Problem 22 at the end of this section.

These observations on consimilarity and condiagonalization help to put into perspective Takagi's factorization (4.4.4) for complex symmetric matrices and Theorem (4.4.3) on triangularization by unitary congruence. Theorem (4.4.3) says that every matrix $A \in M_n$ such that $A\bar{A}$ has all nonnegative eigenvalues can be unitarily contriangularized, and Takagi's result says that every complex symmetric matrix can be unitarily condiagonalized.

Since there is no useful distinction between "real" and "not real" for coneigenvalues, there is no distinction between "unitarily condiagonalizable with real (or positive) coneigenvalues," which would be an analog of Hermitian (or positive definite), and "unitarily condiagonalizable with complex coneigenvalues," which would be an analog of normal. Thus, complex symmetric matrices may be thought of as an analog (for consimilarity) of the whole class of normal matrices (for ordinary similarity), and Takagi's factorization may be thought of as an analog of the spectral theorem for normal matrices (2.5.4a, b).

The theory of ordinary similarity arises as a result of studying linear transformations referred to different bases. In its general context, consimilarity arises as a result of studying antilinear transformations referred to different bases. An antilinear transformation T is a mapping $T: V \to W$ from one complex vector space into another that is additive $[T(x+y) = Tx + Ty$ for all $x, y \in V]$ but conjugate homogeneous $[T(ax) = \bar{a}Tx$ for all $a \in \mathbf{C}$ and all $x \in V$, sometimes called antihomogeneous]. Such transformations occur in quantum mechanics in the study of time reversal.

The class of condiagonalizable matrices is a wide one, which includes all real diagonalizable matrices with real eigenvalues, all (real or complex) symmetric matrices, and all matrices of the form H^2S with H Hermitian and S symmetric (see Problems 10 and 11 at the end of this section). The latter observation is the basis for the second of the following useful sufficient conditions. A *positive definite* matrix $A \in M_n$ is a nonsingular Hermitian matrix such that $x^*Ax > 0$ for all nonzero $x \in \mathbf{C}^n$; an equivalent condition on a Hermitian matrix A is that all the eigenvalues of A are positive or that $A = H^2$ for some nonsingular Hermitian matrix H (see Chapter 7).

4.6.12 Corollary. Let $A, B \in M_n$ with A Hermitian and positive definite.

(a) If B is Hermitian, then there exists a nonsingular $S \in M_n$ such that $SAS^* = I$ and SBS^* is real diagonal.

(b) If B is symmetric, then there exists a nonsingular $S \in M_n$ such that $SAS^* = I$ and SBS^T is real diagonal with nonnegative main diagonal entries.

Proof: Let $A = H^2$, where $H \in M_n$ is a nonsingular Hermitian matrix.

(a) $C \equiv A^{-1}B = H^{-2}B$, so C is similar to $HCH^{-1} = H(H^{-2}B)H^{-1} = H^{-1}BH^{-1}$, which is Hermitian and hence is diagonalizable with real eigenvalues; the matrix C must also be diagonalizable with real eigenvalues. Thus, A and B are simultaneously diagonalizable by *congruence by I(b)(2) of (4.5.15). If $H^{-1}BH^{-1} = U\Lambda U^*$ with U unitary and Λ diagonal, the nonsingular matrix $S = U^*H^{-1}$ will make $SAS^* = I$ and $SBS^* = \Lambda$.

(b) $C \equiv A^{-1}B = H^{-2}B$, so $C\bar{C} = H^{-2}B\bar{H}^{-2}\bar{B}$ is similar to

$$H(C\bar{C})H^{-1} = H^{-1}B\bar{H}^{-2}\bar{B}H^{-1} = (H^{-1}B\bar{H}^{-1})(H^{-1}B\bar{H}^{-1})^*$$

which is Hermitian and positive semidefinite and hence is diagonalizable with nonnegative eigenvalues. But

$$\operatorname{rank}(C\bar{C}) = \operatorname{rank}(H^{-1}B\bar{H}^{-1})(H^{-1}B\bar{H}^{-1})^* = \operatorname{rank}(H^{-1}B\bar{H}^{-1})$$

by (0.4.6d), and $\operatorname{rank}(H^{-1}B\bar{H}^{-1}) = \operatorname{rank}(H^{-2}B) = \operatorname{rank} C$ by (0.4.6b). By (4.6.11), therefore, condition III(b)(1) of (4.5.15) is satisfied and so there must be a nonsingular matrix $S \in M_n$ such that SAS^* and SBS^T are both diagonal. Observe that $HC(H^{-1})^T = H(H^{-2}B)(H^{-1})^T = H^{-1}B(H^{-1})^T$ is symmetric, so by (4.4.4) there is a unitary matrix U and a nonnegative diagonal matrix Σ such that $H^{-1}B(H^{-1})^T = U\Sigma U^T$, or $(U^*H^{-1})B(U^*H^{-1})^T = \Sigma$. If we set $S = U^*H^{-1}$, then $S^*AS = I$ as well. □

We have considered consimilarity to a diagonal matrix, but not every matrix is condiagonalizable, and it is natural to ask whether there is some simple form that any matrix can be reduced to under consimilarity. There is a normal form under consimilarity that plays a role analogous to the Jordan form for ordinary similarity. Using it, one can show that for each $A \in M_n$, A is consimilar to $\bar{A}$, A^*, and A^T [compare with (3.2.3)], A is consimilar to a Hermitian matrix [compare with (4.4.9)], A is consimilar to a real matrix, and there are nonsingular symmetric matrices $S_1, S_2 \in M_n$ and Hermitian matrices $H_1, H_2 \in M_n$ such that $A = S_1H_1 = H_2S_2$ [compare with Corollary (4.4.11)]. In fact, one can reduce the whole question of consimilarity to more familiar notions: Two matrices $A, B \in M_n$ are consimilar if and only if (a) $A\bar{A}$ is similar to $B\bar{B}$, and (b) $\operatorname{rank} A = \operatorname{rank} B$, $\operatorname{rank} A\bar{A} = \operatorname{rank} B\bar{B}$, $\operatorname{rank} A\bar{A}A = \operatorname{rank} B\bar{B}B, \ldots$, and so on for all n such alternating products with at most n terms.

Problems

1. Show that consimilarity is an equivalence relation on M_n.
2. Provide the details for a proof of Theorem (4.6.3).

3. Let $A \in M_n$ be given, and let λ be a coneigenvalue of A. Show that the set of coneigenvectors of A corresponding to λ is not necessarily a subspace of $\mathbf{C}^n$ over $\mathbf{C}$ but is always a subspace over $\mathbf{R}$. Contrast with the situation for the ordinary eigenvectors of A.

4. Theorem (4.6.11) gives necessary and sufficient conditions for a single matrix to be condiagonalizable, but what if one has several matrices that are to be condiagonalized simultaneously? Let $\{A_1, A_2, \ldots, A_k\} \subset M_n$ be given and suppose there is a nonsingular $S \in M_n$ such that $A_i = S\Lambda_i \bar{S}^{-1}$ for $i = 1, \ldots, k$ and each Λ_i is diagonal. Show that (a) each A_i is condiagonalizable; (b) each $A_i\bar{A}_j$ is diagonalizable; (c) the family of products $\{A_i\bar{A}_j : i, j = 1, \ldots, k\}$ commutes; and (d) $A_i\bar{A}_j + A_j\bar{A}_i$ has only real eigenvalues and $A_i\bar{A}_j - A_j\bar{A}_i$ has only imaginary eigenvalues for all $i, j = 1, \ldots, k$. What does this say when $k = 1$? In fact, these necessary conditions are also sufficient; for a proof see the paper of Hong and Horn referenced at the end of Section (4.5).

5. The matrix $A\bar{A}$ plays an important role in the theory of consimilarity. Show that for any $A \in M_n$, the characteristic polynomial of $A\bar{A}$ has real coefficients and conclude that any complex eigenvalues of $A\bar{A}$ must occur in conjugate pairs. *Hint:* $\det(tA - A\bar{A}A) = \det A \det(tI - \bar{A}A) = \det(tI - A\bar{A}) \det A$. Thus, if A is nonsingular, the characteristic polynomials of $\bar{A}A$ and $A\bar{A} = \overline{(\bar{A}A)}$ are the same. Consider $A_\epsilon = A + \epsilon I$ for the general case. See Problem 8 for a more specific result about $A\bar{A}$.

6. The nonnegative eigenvalues of $A\bar{A}$ lead to coneigenvalues of A, but any eigenvalues of $A\bar{A}$ that are not nonnegative also have a significance. Suppose $A \in M_n$ and $A\bar{x} = \lambda x$ for some $x \neq 0$ and $\lambda \in \mathbf{C}$ such that $\lambda \notin [0, \infty)$. Let $\alpha \in \mathbf{C}$ be any square root of λ and define the vector y by $A\bar{x} = \bar{\alpha}y$. Show that $A\bar{y} = \alpha x$, $A\bar{A}y = \bar{\lambda}y$, and the vectors x and y are independent. *Hint:* If they are dependent, x must be a coneigenvector and $\lambda \geq 0$. Conclude that all the complex eigenvalues of $A\bar{A}$ must occur in conjugate pairs and that any negative eigenvalue of $A\bar{A}$ must have geometric multiplicity at least two. Compare with Problem 5.

7. Let $A \in M_n$, and suppose λ is a real strictly negative eigenvalue of $A\bar{A}$, $A\bar{A}x = \lambda x$, $x \neq 0$, $\alpha^2 = \lambda$, $A\bar{x} = \bar{\alpha}y$, $A\bar{y} = \alpha x$. According to Problem 6, x and y are independent. (a) Let $x' = x + \beta y$, $y' = y - \bar{\beta}x$. Show that $A\bar{x}' = \bar{\alpha}y'$ and $A\bar{y}' = \alpha x'$ for any choice of $\beta \in \mathbf{C}$. (b) Show that β can be chosen so that x' and y' are orthogonal, and make such a choice for β. (c) Let $s > 0$ be such that $\xi = sx'$ is a unit vector, and let $\eta = sy'$. Show that $A\bar{\xi} = \bar{\alpha}\eta$, $A\bar{\eta} = \alpha\xi$, and $\xi^*\eta = 0$. (d) Let $r > 0$ be such that $r\eta$ is unit vector and let $U = [\eta \; r\eta \; u_3 \; \cdots \; u_n] \in M_n$ be unitary. Show that

$$U^*A\bar{U} = \left[\begin{array}{cc|c} 0 & r\alpha & \\ \bar{\alpha}/r & 0 & * \\ \hline \multicolumn{2}{c|}{0} & A' \end{array}\right] \quad \text{with} \quad A' \in M_{n-2}$$

and hence that

$$U^*(A\bar{A})U = \left[\begin{array}{cc|c} \lambda & 0 & \\ 0 & \lambda & * \\ \hline \multicolumn{2}{c|}{0} & A'\bar{A}' \end{array}\right]$$

(e) Conclude that every negative eigenvalue of $A\bar{A}$ has even algebraic multiplicity. Compare with Problem 6.

8. For any $A \in M_n$, show that

$$\begin{bmatrix} I & -A \\ 0 & I \end{bmatrix} \begin{bmatrix} A\bar{A} & 0 \\ \bar{A} & 0 \end{bmatrix} \begin{bmatrix} I & A \\ 0 & I \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ \bar{A} & \bar{A}A \end{bmatrix}$$

Conclude from this explicit similarity that there is a one-to-one correspondence between the Jordan blocks of $A\bar{A}$ and $\bar{A}A$ that have nonzero eigenvalues. Since $\bar{A}A = \overline{A\bar{A}}$, show that the Jordan blocks of $A\bar{A}$ with complex eigenvalues occur in conjugate pairs. Conclude that $A\bar{A}$ is similar to a real matrix for any $A \in M_m$. *Hint:* See the discussion of the real Jordan form in (3.4). Somewhat more is actually true. In fact, $A\bar{A}$ is always similar to the square of a real matrix. What does this imply about the eigenvalues of $A\bar{A}$?

9. If $A \in M_n$ is similar to a real matrix, show that A is similar to $\bar{A}$ (and conversely). Use this fact and Problem 8 to show that although AB need not be similar to BA in general, nevertheless $A\bar{A}$ is always similar to $\bar{A}A$ for any $A \in M_n$.

10. Show that the set of condiagonalizable matrices in M_n includes the following: (a) All real diagonalizable matrices with only real eigenvalues. (b) All diagonalizable matrices with a linearly independent set of n *real* eigenvectors. (c) All symmetric matrices. (d) All positive definite Hermitian matrices. *Hint:* $A = HH = H(HH^T)\bar{H}^{-1}$ if A is positive definite; H is Hermitian and nonsingular. (e) All matrices of the form AB, where A is positive definite Hermitian and B is symmetric. This is the same as

the set of all matrices of the form H^2B with H Hermitian nonsingular and B symmetric. *Hint:* $H^2B = H(HBH^T)\bar{H}^{-1}$.

11. Show that the set CD_n of condiagonalizable matrices in M_n has the following properties: (a) If $A \in CD_n$ and $S \in M_n$ is nonsingular, then $SA\bar{S}^{-1} \in CD_n$. (b) The zero matrix is in CD_n. (c) If $A \in CD_n$ and $a \in \mathbf{C}$, then $aA \in CD_n$. (d) If $A \in CD_n$ is invertible, then $A^{-1} \in CD_n$.

12. Show that (a) $\begin{bmatrix} i & 1 \\ 0 & i \end{bmatrix}$ is not diagonalizable in the ordinary sense but it is condiagonalizable. (b) $\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ is diagonalizable in the ordinary sense but is not condiagonalizable. (c) $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ is neither diagonalizable nor condiagonalizable.

13. If $A \in M_n$ is such that $A\bar{A} = \Lambda = \lambda_1 I_{n_1} \oplus \cdots \oplus \lambda_k I_{n_k}$ with $\lambda_i \neq \lambda_j$ if $i \neq j$, and all $\lambda_i \geq 0$, show that there is a unitary $U \in M_n$ such that $A = U\Delta U^T$ and $\Delta = \Delta_1 \oplus \cdots \oplus \Delta_k$, where each $\Delta_i \in M_{n_i}$ is upper triangular.

14. Lemma (4.6.9) says that $A \in M_n$ has the factorization $A = S\bar{S}^{-1}$ for some nonsingular $S \in M_n$ if and only if $A\bar{A} = I$. Use (4.4.4) to show that $A = U\bar{U}^{-1} = UU^T$ for some unitary $U \in M_n$ if and only if $A^{-1} = \bar{A}$ and A is symmetric. What does this have to do with (4.4.7)?

15. Let $A \in M_n$ and write $A = B + iC$ with $B, C \in M_n(\mathbf{R})$. Show that $\lambda \in \mathbf{C}$ is a coneigenvalue of A if and only if $\pm|\lambda|$ are (real) eigenvalues of the block matrix

$$F = \begin{bmatrix} B & C \\ C & -B \end{bmatrix} \in M_{2n}(\mathbf{R})$$

Hint: Write $A\bar{x} = rx$ in terms of $x = u + iv$, $u, v \in \mathbf{R}^n$, $r = |\lambda|$. Thus, if F has no real eigenvalues, A can have no coneigenvalues.

16. Show that if $A \in M_n$ is diagonal or upper triangular, then the eigenvalues of A and the coneigenvalues of A are "the same" in the following sense: If λ is an eigenvalue of A, then $e^{i\theta}\lambda$ is a coneigenvalue of A for *all* $\theta \in \mathbf{R}$, and if μ is a coneigenvalue of A, then $e^{i\theta}\mu$ is an eigenvalue of A for *some* $\theta \in \mathbf{R}$.

17. If $A \in M_n(\mathbf{R})$, show that every real eigenvalue of A is also a coneigenvalue of A and that if $\mu \geq 0$ is a coneigenvalue of A, then either μ or $-\mu$ is an eigenvalue of A. *Hint:* Write $A\bar{x} = \mu x$ in terms of $x = u + iv$, $u, v \in \mathbf{R}^n$. Consider the example $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ to show that a real matrix can have nonreal eigenvalues that are not associated with any coneigenvalues.

18. What does Lemma (4.6.9) say when $n = 1$? A complex number z lies on the unit circle in the complex plane if $z\bar{z} = 1$. The usual generalization of this condition to matrices is to require that $AA^* = I$; such matrices are unitary and play a fundamental role in matrix theory. Another generalization (which reduces to the same thing when $n = 1$) is to require that $A\bar{A} = I$, and it is these matrices that Lemma (4.6.9) characterizes as being consimilar to the identity matrix. Show that if $A = M_n$ and $A\bar{A} = I$, then (a) A is nonsingular; (b) $A^{-1} = \bar{A}$; (c) $|\det A| = |\lambda_1 \cdots \lambda_n| = 1$; and (d) if $Ax = \lambda x$ and $x \neq 0$, then $A\bar{x} = (1/\bar{\lambda})\bar{x}$; so $1/\bar{\lambda}$ is an eigenvalue of A whenever λ is an eigenvalue of A. Show that the matrix $B = \begin{bmatrix} z & i \\ -i & z \end{bmatrix}$, $z \in \mathbf{R}$, $z \neq \pm 1$, has the property that the spectrum of $A = B\bar{B}^{-1}$ is

$$\left\{ \frac{z-1}{z+1}, \frac{z+1}{z-1} \right\}$$

so not all eigenvalues of such matrices lie on the unit circle.

19. It is a fact that every complex matrix $A \in M_n$ can be written as $A = RE$, where $R, E \in M_n$, R is similar to a real matrix, and $E\bar{E} = I$. Show how this decomposition follows from the fact that every $A \in M_n$ is consimilar to a real matrix and explain how it generalizes the fact that every complex number z can be written as $z = re^{i\theta}$ with r and θ real.

20. Show that Theorem (4.6.11) follows from the general necessary and sufficient conditions for two matrices to be consimilar that were stated in the last paragraph of the text of this section. *Hint:* Apply the conditions to A and a diagonal matrix Λ.

21. Use the fact that every $A \in M_n$ is consimilar to a real matrix to show that A must have at least one coneigenvalue if n is odd. *Hint:* A real matrix R of odd order has at least one real eigenvalue. What does this say about the eigenvalues of R^2? If A is consimilar to R, how is $A\bar{A}$ related to R^2?

22. Let $A \in M_n$ be symmetric. The discussion after Theorem (4.6.11) shows that A is condiagonalizable, so there exists a nonsingular $S \in M_n$ and a diagonal $\Lambda \in M_n$ such that $A = S\Lambda\bar{S}^{-1}$. Show that one may take S to be unitary [and hence deduce Corollary (4.4.4) from Theorem (4.6.11)] as follows: Observe that the symmetry of A implies that $(S^*S)\Lambda = \Lambda(\overline{S^*S}) = \Lambda(S^*S)^T$. Use the polar decomposition (7.3.3) to write $S = UP$, where $U \in M_n$ is unitary, $P \in M_n$ is Hermitian, and $P = p(S^*S)$ for some polynomial $p(t)$ [see the proof of Theorem (7.2.6)]. Deduce that $P\Lambda = \Lambda\bar{P} = \Lambda P^T$ and hence $S\Lambda\bar{S}^{-1} = U\Lambda U^T$.

Further Readings. For more information about consimilarity and the problem of simultaneous condiagonalization of a family of matrices, see

the papers of Hong and Horn referenced at the end of Sections (4.4) and (4.5) as well as Y. P. Hong and R. A. Horn, "A Canonical Form for Matrices under Consimilarity," *Linear Algebra Appl.* 102 (1988), 143–168. The notion of consimilarity can be generalized by replacing the complex field with an arbitrary field and replacing the operation of complex conjugation by an automorphism on the field; see [Jac], p. 27.

CHAPTER 5

Norms for vectors and matrices

5.0 Introduction

If one has several vectors in $\mathbf{C}^n$ or several matrices in M_n, what might it mean to say that some are "small" or that others are "large"? Under what circumstances might we say that two vectors are "close together" or "far apart"?

Questions of "size" and "proximity" in a two- or three-dimensional real vector space usually refer to Euclidean distance. The Euclidean length of a vector $z \in \mathbf{R}^n$ is $(z^Tz)^{1/2} = (\sum z_i^2)^{1/2}$, and z is said to be "small" (with respect to this measure) if this nonnegative real number is small. The vectors x and y, furthermore, are "close" if the Euclidean length of the difference $z = x - y$ is a small number.

What may be said about the "size" of matrices, which may be thought of as vectors in a higher-dimensional space? What about vectors in infinite-dimensional spaces? What about complex vectors? Are there useful ways to measure the "size" of real vectors other than by Euclidean length?

One way to answer these questions is to study *norms,* or measures of size, of matrices and vectors. Norms may be thought of as generalizations of Euclidean length, but the study of norms is more than an exercise in mathematical generalization. It is necessary for a proper formulation of notions such as power series of matrices, and it is essential in the analysis and assessment of algorithms for numerical computations. Furthermore, different acceptable norms may be more or less convenient in various situations, so that it is appropriate to study properties common to all norms, rather than to restrict attention to any single norm.

The following examples indicate a few ways in which the need for norms arises.

5.0.1 **Example** (convergence). If x is a complex number such that $|x|<1$, we know that

$$(1-x)^{-1}=1+x+x^2+x^3+\cdots$$

This suggests the formula

$$(I-A)^{-1}=I+A+A^2+A^3+\cdots$$

for calculating the inverse of the square matrix $I-A$, but when is it valid? It turns out that it is sufficient that a matrix norm of A be less than 1, and any such norm will do! Similarly, many other power series which can be used to define matrix-valued functions of a matrix, such as

$$e^A \equiv \sum_{k=0}^{\infty} \frac{1}{k!} A^k$$

can be shown to be convergent and well-defined using norms. Norms may also be useful in determining the number of terms required in a power series in order to calculate a particular function value to a desired degree of accuracy. Similar remarks may be made about the analysis of convergence of iterative schemes to solve systems of equations.

5.0.2 **Example** (accuracy). If f is a real scalar-valued differentiable function of a real variable, we know that if the value of $f(x)$ is known for $x=x_0$, then its value at nearby points $x=x_0+h$ can be estimated in terms of the first derivative

$$\frac{f(x_0+h)-f(x_0)}{h}=\frac{\Delta f}{\Delta x}\cong f'(x_0)$$

Thus, we have a way of estimating the relative error in computing the value of f at x_0 if we actually compute the value of f at a nearby point x_0+h instead.

The same issue arises for matrix calculations. Suppose we wish to compute A^{-1} (or some other function of A), but the entries of A are obtained by experiment, by analysis of other data, or from prior calculation, and they are not known exactly. We may think of A as being composed of the "true" A_0 plus an error E, and we would like to assess the potential "relative error" (in terms of the "size" of E) in computing $A^{-1}=(A_0+E)^{-1}$ instead of the true A_0^{-1}. Bounds for the disparity

between A^{-1} and A_0^{-1} may be as important to know as the exact value of A^{-1}, and norms provide a systematic way of dealing with such questions.

5.0.3 **Example** (bounds). Bounds for important quantities associated with a matrix, such as eigenvalues, often involve norms, as do bounds for possible changes in these quantities when a matrix is perturbed.

5.1 Defining properties of vector norms and inner products

We first consider norms on a vector space. Since M_n is a vector space, everything we do will also apply to norms of matrices.

In order to specify properties to be required of a function if it is to be a norm, we abstract from the familiar notion of absolute value of (real or complex) scalars. Of course, a significant difference is that, while the absolute value function is a real-valued function of one real or complex variable, we require a norm to be a real-valued function of the several variables that describe a vector. One such function on $\mathbf{C}^n$ is Euclidean length $(z^*z)^{1/2}$, but there are other functions that share some fundamental properties of Euclidean length and may be more relevant measures in some instances, may impart additional information, or may be more convenient to use in certain contexts.

Throughout this chapter we shall consider real or complex vector spaces only. All of the major results hold for both fields, but within each result one must be consistent as to which field is used. Thus, we shall often state results in terms of a field $\mathbf{F}$ (with $\mathbf{F}=\mathbf{R}$ or $\mathbf{C}$ at the outset) and then refer to the same field $\mathbf{F}$ in the rest of the argument.

5.1.1 **Definition.** Let V be a vector space over a field $\mathbf{F}$ ($\mathbf{R}$ or $\mathbf{C}$). A function $\|\cdot\|: V\to\mathbf{R}$ is a *vector norm* if for all $x, y\in V$,

(1)	$\|x\|\geq 0$	Nonnegative
(1a)	$\|x\|=0$ if and only if $x=0$	Positive
(2)	$\|cx\|=|c|\,\|x\|$ for all scalars $c\in\mathbf{F}$	Homogeneous
(3)	$\|x+y\|\leq\|x\|+\|y\|$	Triangle inequality

These four axioms are familiar properties of Euclidean length in the plane. Euclidean length possesses other properties that are independent of these four axioms [e.g., the parallelogram identity (5.1.8)], which we do not adopt as axioms because they are not essential to the general theory.

A function that satisfies axioms (1), (2), and (3), but not necessarily (1a) is called a *vector seminorm*. A seminorm generalizes the notion of a

norm in that some vectors other than the zero vector are allowed to have zero length.

5.1.2 **Lemma.** If $\|\cdot\|$ is a vector seminorm on V, then

$$|\,\|x\|-\|y\|\,|\leq\|x-y\|$$

for all $x, y\in V$.

Proof: Since $y=x+(y-x)$, we have

$$\|y\|\leq\|x\|+\|y-x\|=\|x\|+\|x-y\|$$

from the triangle inequality (3) and the homogeneity axiom (2). From this it follows that

$$\|y\|-\|x\|\leq\|x-y\|$$

But $x=y+(x-y)$ as well, so we have

$$\|x\|\leq\|y\|+\|x-y\|$$

from the triangle inequality (3) again, and hence

$$\|x\|-\|y\|\leq\|x-y\|$$

Thus, we have shown that $\pm(\|x\|-\|y\|)\leq\|x-y\|$, which is equivalent to the assertion of the lemma. □

Associated with Euclidean length on $\mathbf{C}^n$ is the usual Euclidean inner product y^*x (sometimes called the "dot product"), which has something to do with the "angle" between two vectors: x and y are orthogonal if $y^*x=0$. Just as for vector norms, one can abstract a few essential characteristics of the Euclidean inner product and use them as axioms for a general theory of inner products.

5.1.3 **Definition.** Let V be a vector space over the field $\mathbf{F}$ ($\mathbf{R}$ or $\mathbf{C}$). A function $\langle\cdot,\cdot\rangle: V\times V\rightarrow\mathbf{F}$ is an *inner product* if for all $x, y, z\in V$,

(1)	$\langle x,x\rangle\geq 0$	Nonnegative
(1a)	$\langle x,x\rangle=0$ if and only if $x=0$	Positive
(2)	$\langle x+y,z\rangle=\langle x,z\rangle+\langle y,z\rangle$	Additive
(3)	$\langle cx,y\rangle=c\langle x,y\rangle$ for all scalars $c\in\mathbf{F}$	Homogeneous
(4)	$\langle x,y\rangle=\overline{\langle y,x\rangle}$	Hermitian property

Exercise. Show that the Euclidean inner product $\langle x,y\rangle=y^*x$ satisfies all four of the above axioms for an inner product.

Exercise. Let $D=\mathrm{diag}(d_1, d_2, \ldots, d_n)$ and consider the function $(x,y)\equiv y^*Dx$. Which of the axioms for an inner product does $(\cdot,\cdot)$ satisfy? Under what conditions on D is $(\cdot,\cdot)$ an inner product?

Exercise. Deduce the following properties of an inner product from the four axioms in Definition (5.1.3):

(a) $\langle x,cy\rangle=\bar{c}\langle x,y\rangle$
(b) $\langle x,y+z\rangle=\langle x,y\rangle+\langle x,z\rangle$
(c) $\langle ax+by,cw+dz\rangle=a\bar{c}\langle x,w\rangle+b\bar{c}\langle y,w\rangle+a\bar{d}\langle x,z\rangle+b\bar{d}\langle y,z\rangle$
(d) $\langle x,y\rangle=0$ for all $y\in V$ if and only if $x=0$
(e) $\langle x,\langle x,y\rangle y\rangle=|\langle x,y\rangle|^2$

An important property shared by all inner products is the Cauchy–Schwarz inequality.

5.1.4 **Theorem** (Cauchy–Schwarz inequality). If $\langle\cdot,\cdot\rangle$ is an inner product on a vector space V over the field $\mathbf{F}$ ($\mathbf{R}$ or $\mathbf{C}$), then

$$|\langle x,y\rangle|^2\leq\langle x,x\rangle\langle y,y\rangle \quad\text{for all}\quad x,y\in V$$

Equality occurs if and only if x and y are linearly dependent, that is, $x=\alpha y$ or $y=\alpha x$ for some $\alpha\in\mathbf{F}$.

Proof: Let $x, y\in V$ be given. If $y=0$, the assertion is trivial, so we may assume that $y\neq 0$. Let $t\in\mathbf{R}$ and consider $p(t)\equiv\langle x+ty,x+ty\rangle=\langle x,x\rangle+t\langle y,x\rangle+t\langle x,y\rangle+t^2\langle y,y\rangle=\langle x,x\rangle+2t\,\mathrm{Re}\langle x,y\rangle+t^2\langle y,y\rangle$, which is a real quadratic polynomial with real coefficients. Because of axiom (5.1.3(1)), we know that $p(t)\geq 0$ for all real t, and hence $p(t)$ can have no real simple roots. The discriminant of $p(t)$ must therefore be nonpositive

$$(2\,\mathrm{Re}\langle x,y\rangle)^2-4\langle y,y\rangle\langle x,x\rangle\leq 0$$

and hence

$$(\mathrm{Re}\langle x,y\rangle)^2\leq\langle x,x\rangle\langle y,y\rangle \tag{5.1.5}$$

Since this inequality must hold for any pair of vectors, it must hold if y is replaced by $\langle x,y\rangle y$, so we also have the inequality

$$(\mathrm{Re}\langle x,\langle x,y\rangle y\rangle)^2\leq\langle x,x\rangle\langle y,y\rangle|\langle x,y\rangle|^2$$

But $\mathrm{Re}\langle x,\langle x,y\rangle y\rangle=\mathrm{Re}\overline{\langle x,y\rangle}\langle x,y\rangle=\mathrm{Re}|\langle x,y\rangle|^2=|\langle x,y\rangle|^2$, so

$$|\langle x,y\rangle|^4\leq\langle x,x\rangle\langle y,y\rangle|\langle x,y\rangle|^2 \tag{5.1.6}$$

If $\langle x,y\rangle=0$, then the statement of the theorem is trivial; if not, then we may divide (5.1.6) by the quantity $|\langle x,y\rangle|^2$ to obtain the desired

inequality. Because of axiom (1a), $p(t)$ can have a real (double) root only if $x+ty=0$ for some t. Thus, equality can occur in the discriminant condition (5.1.5) if and only if x and y are linearly dependent. $\square$

5.1.7 **Corollary.** If $\langle\bullet,\bullet\rangle$ is a vector inner product on V, then $\|x\|\equiv(\langle x,x\rangle)^{1/2}$ is a vector norm on V.

Exercise. Prove (5.1.7). *Hint:* The only nontrivial axiom to verify is the triangle inequality. Compute $\|x+y\|^2$ and use the Cauchy-Schwarz inequality.

If $\|\bullet\|$ is a vector norm such that $\|x\|=\langle x,x\rangle^{1/2}$ for some inner product $\langle\bullet,\bullet\rangle$, then we say that the vector norm $\|\bullet\|$ is *derived from an inner product* (namely, from $\langle\bullet,\bullet\rangle$).

Problems

1. Let e_i denote the ith unit coordinate vector in $\mathbf{C}^n$ and suppose that $\|\bullet\|$ is a vector seminorm on $\mathbf{C}^n$. Show that

$$\|x\|\le\sum_{i=1}^{n}|x_i|\,\|e_i\|$$

2. If $\|\bullet\|$ is a vector seminorm on V, show that $V_0=\{v\in V\colon \|v\|=0\}$ is a subspace of V (called the *null space* of $\|\bullet\|$). (a) If V_1 is any subspace of V such that $V_0\cap V_1=\{0\}$, show that $\|\bullet\|$ is a vector norm on V_1. (b) Consider the relation $x\sim y$ defined by

$$x\sim y \quad\text{if and only if}\quad \|x-y\|=0$$

Show that $\sim$ is an equivalence relation on V, that the cosets of this equivalence relation are of the form $\hat{x}=\{x+y\in V\colon y\in V_0\}$, and that the set of these cosets forms a vector space in a natural way. Show that the function $\|\hat{x}\|\equiv\{\|x\|\colon x\in\hat{x}\}$ is well defined and is a vector norm on the vector space of cosets. (c) Explain why there is a natural norm associated with every vector seminorm. (d) Is $\|x\|\equiv 0$ a seminorm? (e) Give an example of a nontrivial seminorm that is not a norm.

3. Show that if we define the "angle" between the nonzero vectors x and y to be the value of

$$\cos^{-1}\left(\frac{|\langle x,y\rangle|}{(\langle x,x\rangle\langle y,y\rangle)^{1/2}}\right)$$

that lies between 0 and $\pi/2$, then this notion of angle is well defined for any inner product.

4. Show that any vector norm derived from an inner product as in (5.1.7) must satisfy the *parallelogram identity*

$$\tfrac{1}{2}(\|x+y\|^2+\|x-y\|^2)=\|x\|^2+\|y\|^2 \tag{5.1.8}$$

Why is this identity so named? The equation (5.1.8) is, in fact, necessary and sufficient that a given norm $\|\bullet\|$ be derived from an inner product. See Problem 10.

5. Consider the function $\|x\|_\infty\equiv\max_{1\le i\le n}|x_i|$ defined on $\mathbf{C}^n$. Show that $\|\bullet\|_\infty$ is a vector norm that cannot be derived from an inner product.

6. If $\|\bullet\|$ is a vector norm derived from an inner product $\langle\bullet,\bullet\rangle$, show that

$$\operatorname{Re}\langle x,y\rangle=\tfrac{1}{4}(\|x+y\|^2-\|x-y\|^2) \tag{5.1.9}$$

This is known as the *polarization identity.* Show also that

$$\operatorname{Re}\langle x,y\rangle=\tfrac{1}{2}(\|x+y\|^2-\|x\|^2-\|y\|^2)$$

7. Show that the l_1 norm $\|x\|\equiv|x_1|+\cdots+|x_n|$ on $\mathbf{C}^n$ satisfies the axioms (5.1.1) but does not obey the polarization identity (5.1.9). It is not, therefore, derived from any inner product.

8. If $\|\bullet\|$ is a vector norm on V derived from an inner product, then

$$\|x+y\|\,\|x-y\|\le\|x\|^2+\|y\|^2$$

for all $x,y\in V$. When does equality hold? Does this inequality hold for all vector norms? Give a geometric interpretation of this inequality.

9. Let x and y be given vectors in V, which has a norm $\|\bullet\|$ derived from an inner product $\langle\bullet,\bullet\rangle$, and suppose that y is nonzero. Show that the scalar α_0 that minimizes the value of $\|x-\alpha y\|$ is $\alpha_0=\langle x,y\rangle/\|y\|^2$, and that $x-\alpha_0 y$ and y are othogonal.

10. It is not difficult to show that the parallelogram identity (5.1.8) is a sufficient condition for a given norm to be derived from an inner product, but some ingenuity is required. First consider the case of a vector space V over $\mathbf{R}$. Let $\|\bullet\|$ be a given norm on V. (a) Define

$$\langle x,y\rangle=\frac{\|x+y\|^2-\|x\|^2-\|y\|^2}{2} \tag{5.1.10}$$

Show that $\langle\bullet,\bullet\rangle$ defined in this way satisfies axioms (1), (1a), and (4) in (5.1.3) and that $\langle x,x\rangle=\|x\|^2$. (b) Use (5.1.8) to show that

$$\begin{aligned}4\langle x,y\rangle+4\langle z,y\rangle&=2\|x+y\|^2+2\|z+y\|^2-2\|x\|^2-2\|z\|^2-4\|y\|^2\\&=\|x+2y+z\|^2-\|x+z\|^2-4\|y\|^2=4\langle x+z,y\rangle\end{aligned}$$

and conclude that the additivity axiom (2) in (5.1.3) is satisfied. (c) Use the additivity axiom to show that $\langle nx,y\rangle=n\langle x,y\rangle$ and $m\langle m^{-1}nx,y\rangle=$

$\langle nx, y\rangle = n\langle x, y\rangle$ whenever m and n are nonnegative integers. Use (5.1.8) and (5.1.10) to show that $\langle -x, y\rangle = -\langle x, y\rangle$ and conclude that $\langle ax, y\rangle = a\langle x, y\rangle$ whenever $a \in \mathbf{R}$ is rational. (d) Let $p(t) = t^2\|x\|^2 + 2t\langle x, y\rangle + \|y\|^2$, $t \in \mathbf{R}$, and show that $p(t) = \|tx+y\|^2$ if t is rational. Conclude from the continuity of $p(t)$ that $p(t) \geq 0$ for all $t \in \mathbf{R}$. Deduce the Cauchy-Schwarz inequality $|\langle x, y\rangle|^2 \leq \|x\|^2\|y\|^2$ from the fact that the discriminant of $p(t)$ must be nonpositive. (e) Now let $a \in \mathbf{R}$ be given. Show that

$$|\langle ax, y\rangle - a\langle x, y\rangle| = |\langle (a-b)x, y\rangle + (b-a)\langle x, y\rangle|$$

$$\leq |\langle (a-b)x, y\rangle| + |(b-a)\langle x, y\rangle| \leq 2|a-b|\,\|x\|\,\|y\|$$

for any rational b, and observe that the upper bound can be made arbitrarily small. Conclude that the homogeneity axiom (3) in (5.1.3) is satisfied. This shows that $\langle \cdot, \cdot \rangle$ is an inner product on V.

A careful reader will observe that the triangle inequality for the norm $\|\cdot\|$ [axiom (3) in (5.1.1)] is not used in this argument. Thus, the axioms (1), (1a), and (2) in (5.1.1) together with (5.1.8) imply that the function $\|\cdot\|$ is derived from an inner product, is therefore a norm, and hence must satisfy the triangle inequality. (f) If V is a complex vector space, define

$$\langle x, y\rangle = \frac{\|x+y\|^2 - \|x\|^2 - \|y\|^2}{2} + \frac{i(\|x+iy\|^2 - \|x\|^2 - \|y\|^2)}{2}$$

The real part of $\langle x, y\rangle$ is an inner product of V considered as a vector space over $\mathbf{R}$. Use this fact and (5.1.8) to show that $\langle \cdot, \cdot \rangle$ is an inner product for V as a vector space over $\mathbf{C}$.

Further Reading. The first proof that that parallelogram identity is both necessary and sufficient for a given vector norm to be derived from an inner product seems to be due to P. Jordan and J. Von Neumann, "On Inner Products in Linear Metric Spaces," *Ann. Math.* 36(2) (1935), 719–723. The outline of a proof of this result given in Problem 10 follows D. Fearnley-Sander and J. S. V. Symons, "Apollonius and Inner Products," *Amer. Math. Monthly* 81 (1974), 990–993.

5.2 Examples of vector norms

The following are some examples of frequently encountered vector norms.

5.2.1 The *Euclidean norm* (or l_2 norm) on $\mathbf{C}^n$ is

$$\|x\|_2 \equiv (|x_1|^2 + \cdots + |x_n|^2)^{1/2}$$

This is perhaps the best known vector norm since $\|x-y\|_2$ measures the standard Euclidean distance between two points $x, y \in \mathbf{C}^n$. This norm is also derived from the usual Euclidean inner product; that is, $\|x\|_2^2 = \langle x, x\rangle = x^*x$.

Exercise. Verify that $\|\cdot\|_2$ is a vector norm on $\mathbf{C}^n$.

Exercise. A norm $\|\cdot\|$ is said to be *unitarily invariant* if $\|Ux\| = \|x\|$ for all $x \in \mathbf{C}^n$ and all unitary matrices $U \in M_n$. Show that the Euclidean norm $\|\cdot\|_2$ is unitarily invariant.

5.2.2 The *sum norm* (or l_1 norm) on $\mathbf{C}^n$ is

$$\|x\|_1 \equiv |x_1| + \cdots + |x_n|$$

This norm is also called the one-norm or, more picturesquely, the *Manhattan norm* because of the rectilinear measurement of length in coordinate directions only.

Exercise. Verify that the sum norm is a vector norm on $\mathbf{C}^n$, but that it is not derived from an inner product. *Hint:* Use (5.1.8).

5.2.3 The *max norm* (or l_∞ norm) on $\mathbf{C}^n$ is

$$\|x\|_\infty \equiv \max\{|x_1|, \ldots, |x_n|\}$$

Exercise. Verify that $\|\cdot\|_\infty$ is a vector norm on $\mathbf{C}^n$.

Exercise. Is $\|\cdot\|_\infty$ derived from an inner product?

5.2.4 The l_p *norm* on $\mathbf{C}^n$ is

$$\|x\|_p \equiv \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$$

for $p \geq 1$.

Exercise. Verify that each l_p norm for $p \geq 1$ is a vector norm on $\mathbf{C}^n$ and that $\|x\|_\infty = \lim_{p\to\infty}\|x\|_p$ for each $x \in \mathbf{C}^n$. *Hint:* The triangle inequality is the only nontrivial axiom to verify. The triangle inequality for the l_p norms is a classical inequality known as *Minkowski's inequality*.

Exercise. Give an example of a vector norm that is not an l_p norm.

The foregoing examples of vector norms have all been norms on $\mathbf{C}^n$, but they can be used to create vector norms on any finite-dimensional

real or complex vector space V. If $\mathcal{B}=\{b^{(1)},\ldots,b^{(n)}\}$ is a basis for V, then recall that

$$x\to[x]_{\mathcal{B}}\equiv\begin{bmatrix}x_1\\ \vdots\\ x_n\end{bmatrix}\in\mathbf{C}^n,\qquad x=\sum_{i=1}^n x_i b^{(i)}$$

is an isomorphism of V onto $\mathbf{C}^n$. If $\|\cdot\|$ is any vector norm on $\mathbf{C}^n$, then

$$\|x\|_{\mathcal{B}}\equiv\|[x]_{\mathcal{B}}\|=\|[x_1,\ldots,x_n]^T\|,\qquad x=\sum_{i=1}^n x_i b^{(i)}$$

is easily shown to be a vector norm on V.

Exercise. Verify the last assertion.

A matrix $B\in M_n$ is said to be an *isometry for the vector norm* $\|\cdot\|$ on $\mathbf{C}^n$ if

$$\|Bx\|=\|x\|\quad\text{for all}\quad x\in\mathbf{C}^n$$

Exercise. Show that an isometry for any vector norm must be a nonsingular matrix.

Exercise. Show that the set of isometries for a given norm forms a group (known as the *isometry group* of the norm). Are there any isometries for $\|\cdot\|_2$ besides the unitary matrices?

Exercise. Show that the isometry group of the sum norm is the set (group) of all matrices that look like permutation matrices except that the "+1" entries are replaced by arbitrary complex numbers with absolute value 1.

Exercise. What is the isometry group of the max norm?

The definition of a vector norm does not require that the vector space V be finite-dimensional. The space V might, for example, be the vector space $C[a,b]$ of all continuous real- or complex-valued functions on the real interval $[a,b]$.

5.2.5 **Example.** Some examples of norms on $C[a,b]$ are similar to norms already defined for C^n. For example,

$$\|f\|_2\equiv\left[\int_a^b|f(t)|^2\,dt\right]^{1/2}\qquad L_2\text{ norm}$$

$$\|f\|_1\equiv\int_a^b|f(t)|\,dt\qquad L_1\text{ norm}$$

$$\|f\|_p\equiv\left[\int_a^b|f(t)|^p\,dt\right]^{1/p},\qquad p\ge 1\qquad L_p\text{ norm}$$

$$\|f\|_\infty\equiv\max\{|f(x)|:x\in[a,b]\}\qquad L_\infty\text{ norm}$$

are all norms on $C[a,b]$.

Problems

1. Show that if $0<p<1$, then (5.2.4) defines a function on $\mathbf{C}^n$ that satisfies all but one of the axioms for a vector norm. Which one fails? Give an example.

2. Let $f\in C[0,1]$. Show that $\|f\|_\infty=\lim_{p\to\infty}\|f\|_p$.

3. What does the triangle inequality look like for $\|\cdot\|_p$ on $C[0,1]$? How could you prove it starting from Minkowski's inequality (Appendix B) for $\mathbf{C}^n$?

4. Let $p_1,p_2,\ldots,p_n$ be given positive real numbers. Which of the following is a vector norm on $\mathbf{C}^n$?

(a) $\|x\|=\sum_{i=1}^n p_i|x_i|$

(b) $\|x\|=\left(\sum_{i=1}^n p_i|x_i|^2\right)^{1/2}$

(c) $\|x\|=\max\{p_1|x_1|,\ldots,p_n|x_n|\}$

5. Let $x_0\in[a,b]$ be a given point. Show that the function $\|f\|_{x_0}\equiv|f(x_0)|$ on $C[a,b]$ is a seminorm that is not a norm if $a<b$.

6. If $\|\cdot\|$ is an unitarily invariant vector norm on $\mathbf{C}^n$, show that $\|\cdot\|=\alpha\|\cdot\|_2$ for some $\alpha>0$ and that $\|\cdot\|_2$ is the only unitarily invariant vector norm for which $\|e_1\|=1$.

7. Show that $\|y\|_\infty=\max_{\|x\|_1=1}|y^*x|$ and that $\|x\|_1=\max_{\|y\|_\infty=1}|x^*y|$.

8. Use the preceding exercise to show that if A^* is in the isometry group of the sum norm, then A is in the isometry group of the max norm, and vice versa.

9. What is the intersection of all the isometry groups of all the l_p norms?

Further Readings. For a detailed discussion of the classical inequalities of Minkowski and Hölder, see [BB].

5.3 Algebraic properties of vector norms

From any given norm or norms, new norms may be constructed in several ways. For example, it is easy to show that the sum of two vector (semi)norms is a vector (semi)norm and any positive multiple of a vector (semi)norm is again a vector (semi)norm. In a different vein, one can also show easily that if $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are vector norms, then the function $\|\cdot\|$ defined by $\|x\| \equiv \max\{\|x\|_\alpha, \|x\|_\beta\}$ is also a vector norm. These observations are all special cases of the following result.

5.3.1 **Theorem.** Let $\|\cdot\|_{\alpha_1}, \ldots, \|\cdot\|_{\alpha_m}$ be m given vector norms on a vector space V over the field $\mathbf{F}$ ($\mathbf{R}$ or $\mathbf{C}$), and let $\|\cdot\|_\beta$ be a vector norm on $\mathbf{R}^m$ such that $\|y\|_\beta \le \|y+z\|_\beta$ for all vectors $y, z \in \mathbf{R}^m$ with nonnegative entries. Then the function $\|\cdot\|: V \to \mathbf{R}$ defined by $\|x\| \equiv \|[\|x\|_{\alpha_1}, \ldots, \|x\|_{\alpha_m}]^T\|_\beta$ is a vector norm on V.

The monotonicity assumption on the norm $\|\cdot\|_\beta$ in the theorem ensures that the constructed function $\|\cdot\|$ satisfies the triangle inequality. All the l_p norms have this monotonicity property, as does any vector norm $\|x\|_\beta$ on $\mathbf{R}^m$ that is a function only of the absolute values of the entries of x; see (5.5.9–10). There are, however, vector norms that do not have this property.

Exercise. Prove Theorem (5.3.1).

Exercise. Show that the fact that the sum or max of two vector norms is a vector norm is a special case of (5.3.1). What about the min?

Exercise. Let $m = 2$, $V = \mathbf{R}^2$, and $\|x\|_\beta = |x_1 - x_2| + |x_2|$. Show that $\|\cdot\|_\beta$ is a vector norm on $\mathbf{R}^2$ but the function $\|x\| \equiv \|[\|x\|_\infty, \|x\|_1]^T\|_\beta = \min\{|x_1|, |x_2|\} + |x_1| + |x_2|$ is not a vector norm. Which of the vector norm axioms does $\|\cdot\|$ satisfy? *Hint:* Consider $x = \begin{bmatrix}1\\0\end{bmatrix}$, $y = \begin{bmatrix}0\\1\end{bmatrix}$, $\|x+y\|$, $\|x\|$, and $\|y\|$. Does this contradict (5.3.1)?

Another way to construct new norms from old is given by the following result.

5.3.2 **Theorem.** If $\|\cdot\|$ is a vector norm on $\mathbf{C}^n$ and if $T \in M_n$ is nonsingular, then $\|\cdot\|_T$ defined by $\|x\|_T \equiv \|Tx\|$, $x \in \mathbf{C}^n$, is also a vector norm on $\mathbf{C}^n$.

Exercise. Prove Theorem (5.3.2).

Exercise. What happens in (5.3.2) if T is singular?

Exercise. Why must $\|x\| \equiv (|2x_1 - 3x_2|^2 + |x_2|^2)^{1/2}$ be a norm on $\mathbf{C}^2$ (no computations, please!).

New norms can be constructed from old ones by using the notion of duality. This method is discussed at the end of Section (5.4).

Problems

1. If $\|\cdot\|$ is a vector seminorm, show that $\|x\|_T \equiv \|Tx\|$ is also a vector seminorm for any $T \in M_n$. If $\|\cdot\|$ is actually a vector norm, then $\|\cdot\|_T$ is a vector seminorm whose null space is the null space of T.

2. Show that any vector seminorm is of the form $\|\cdot\|_T$ for some vector norm $\|\cdot\|$ and some $T \in M_n$.

5.4 Analytic properties of vector norms

The examples in the preceding two sections make it clear that there are many different functions $\|\cdot\|: V \to \mathbf{R}$ that satisfy the axioms for a norm. It is useful to have many different norms available because one norm may be more convenient or more appropriate than another for a given purpose. For example, the l_2 norm is often convenient to use in optimization problems because it is continuously differentiable (except at the origin). On the other hand, the l_1 norm, while differentiable on a smaller set, is popular in statistics because it leads to estimators that can be more robust than the classical regression estimators. The l_∞ norm is often the most natural one to use, since it directly monitors element-by-element convergence, but, unfortunately, it can be analytically and algebraically awkward to use. In actual applications, the norm on which theory is most naturally based and the norm that is most easily calculated in a given situation may not coincide. It is important, therefore, to know what relationship there may be between two different norms. Fortunately, in the finite-dimensional case all norms are "equivalent" in a certain strong sense.

A basic notion in analysis is that of *convergence of a sequence*, and vector norms can be used to measure convergence of a sequence of vectors.

5.4.1 **Definition.** Let V be a vector space over $\mathbf{R}$ or $\mathbf{C}$ and let $\|\cdot\|$ be a norm on V. We say that the sequence $\{x^{(k)}\}$ of vectors in V *converges* to a vector $x \in V$ with respect to the norm $\|\cdot\|$ if and only if $\|x^{(k)} - x\| \to 0$ as $k \to \infty$.

If $\{x^{(k)}\}$ converges to x with respect to the norm $\|\cdot\|$, we write

$$x^{(k)} \underset{\|\cdot\|}{\rightarrow} x \quad \text{or} \quad \lim_{k\to\infty} x^{(k)} = x \quad \text{with respect to} \quad \|\cdot\|$$

It must be made clear which norm is involved in the convergence in question; the issue arises as to whether a given sequence of vectors can converge with respect to one norm but not with respect to another. This ambiguity can happen in an infinite-dimensional vector space.

5.4.2 Example. Consider the sequence $\{f_k\}$ of functions in $C[0,1]$ (the vector space of all real-valued or complex-valued continuous functions on $[0,1]$) defined by

$$\begin{aligned} f_k(x) &= 0 & 0 \le x \le \frac{1}{k} \\ f_k(x) &= 2(k^{3/2}x - k^{1/2}), & \frac{1}{k} \le x \le \frac{3}{2k} \\ f_k(x) &= 2(-k^{3/2}x + 2k^{1/2}), & \frac{3}{2k} \le x \le \frac{2}{k} \\ f_k(x) &= 0, & \frac{2}{k} \le x \le 1 \end{aligned}$$

for $k = 2, 3, 4, \ldots$. One may then calculate that

$$\|f_k\|_1 = \tfrac{1}{2}k^{-1/2} \to 0 \quad \text{as} \quad k \to \infty$$

$$\|f_k\|_2 = \frac{1}{\sqrt{3}} \quad \text{for all} \quad k$$

$$\|f_k\|_\infty = k^{1/2} \to \infty \quad \text{as} \quad k \to \infty$$

Thus, $\lim_{k\to\infty} f_k = 0$ with respect to the L_1 norm but not with respect to the other two norms.

Exercise. Sketch the functions described in the preceding example and verify the assertions made about the L_1, L_2, and L_∞ norms.

Exercise. If $\|\cdot\|$ is a vector norm, if $x^{(k)} \underset{\|\cdot\|}{\rightarrow} x$, and if $x^{(k)} \underset{\|\cdot\|}{\rightarrow} y$, use the triangle inequality to show that $x = y$. Thus it makes sense to talk about *the* limit of a sequence (if any) with respect to a given norm.

Fortunately, the phenomenon in Example (5.4.2) cannot occur in the case of a finite-dimensional vector space. In order to see this, we need a general lemma about the continuity properties of norms.

5.4.3 Lemma. Let $\|\cdot\|$ be a norm on a vector space V over the field $\mathbf{F}$ ($\mathbf{R}$ or $\mathbf{C}$), and let $x^{(1)}, x^{(2)}, \ldots, x^{(m)} \in V$ be given vectors. The function $g: \mathbf{F}^m \to \mathbf{R}$ defined by

$$g(z_1, z_2, \ldots, z_m) \equiv \|z_1x^{(1)} + z_2x^{(2)} + \cdots + z_mx^{(m)}\|$$

is a uniformly continuous function.

Proof: Let $u = \sum_{i=1}^m u_ix^{(i)}$ and $v = \sum_{i=1}^m v_ix^{(i)}$, and calculate

$$\begin{aligned} |g(u_1, \ldots, u_m) - g(v_1, \ldots, v_m)| &= |\,\|u\| - \|v\|\,| \le \|u - v\| \\ &= \left\| \sum_{i=1}^m (u_i - v_i)x^{(i)} \right\| \le \sum_{i=1}^m |u_i - v_i|\,\|x^{(i)}\| \le C \max_{1\le i\le m} |u_i - v_i| \end{aligned}$$

where $C \equiv m\max_{1\le i\le m}\|x^{(i)}\|$. The first inequality comes from Lemma (5.1.2). Notice that the finite constant C depends only upon the norm $\|\cdot\|$ and the m vectors $x^{(1)}, \ldots, x^{(m)}$. If the vectors $x^{(i)}$ are all the zero vector, there is nothing to show, and, if not, then $C > 0$. In order to have $|g(u_1, \ldots, u_m) - g(v_1, \ldots, v_m)| < \epsilon$, we need only choose $|u_i - v_i| < \epsilon/C$. □

Although the vector space V need not be finite-dimensional for the lemma, it is important that the number of vectors $x^{(i)}$ be finite.

Exercise. Deduce from the lemma that every vector norm on $\mathbf{R}^n$ or $\mathbf{C}^n$ is a uniformly continuous function.

Finite dimensionality of V is, however, essential for the following fundamental fact.

5.4.4 Theorem. Let f_1 and f_2 be two real-valued functions on a finite-dimensional vector space V over the field $\mathbf{F}$ ($\mathbf{R}$ or $\mathbf{C}$), and let $\mathcal{B} = \{x^{(1)}, \ldots, x^{(n)}\}$ be a basis for V. Assume that f_1 and f_2 are

(a) Positive: $f_i(x) \ge 0$ for all $x \in V$, $f_i(x) = 0$ if and only if $x = 0$;
(b) Homogeneous: $f_i(\alpha x) = |\alpha| f_i(x)$ for all $\alpha \in \mathbf{F}$ and all $x \in V$; and
(c) Continuous: $f_i(x(z))$ is continuous on $\mathbf{F}^n$, where

$$z = [z_1, \ldots, z_n]^T \in \mathbf{F}^n \quad \text{and} \quad x(z) \equiv z_1x^{(1)} + \cdots + z_nx^{(n)}$$

Then there exist finite positive constants C_m and C_M such that

$$C_m f_1(x) \le f_2(x) \le C_M f_1(x)$$

for all $x \in V$.

Proof: Define $h(z) \equiv f_2(x(z))/f_1(x(z))$ on the Euclidean unit sphere $S = \{z \in \mathbf{F}^n : \|z\|_2 = 1\}$, a compact set in $\mathbf{F}^n$. Notice that the denominator of

$h(z)$ does not vanish on S by (a), and therefore $h(z)$ is continuous on S by (c). By the Weierstrass theorem (see Appendix E), the continuous function h achieves a finite positive maximum C_M and a positive minimum C_m on the compact set S and hence

$$C_m f_1(x(z)) \leq f_2(x(z)) \leq C_M f_1(x(z))$$

for all $z \in S$. Because $z/\|z\|_2 \in S$ for every nonzero $z \in \mathbf{F}^n$, (b) ensures that these inequalities hold for all nonzero $z \in \mathbf{F}^n$; they hold trivially for $z=0$ since $f_i(0)=0$. But every $x \in V$ is of the form $x=x(z)$ for some $z \in \mathbf{F}^n$ because $\mathcal{B}$ is a basis, so the asserted inequalities hold for all $x \in V$. □

Definition. Let V be a real or complex vector space. A function $f: V \to \mathbf{R}$ that satisfies the three hypotheses of positivity, homogeneity, and continuity in Theorem (5.4.4) is said to be a *pre-norm*.

The most important example of a class of pre-norms is, of course, the vector norms; Lemma (5.4.3) says that every vector norm satisfies the continuity assumption (c) of Theorem (5.4.4). A pre-norm that satisfies the triangle inequality is a vector norm. Because of the importance of this class, we state the result in this case as the following corollary.

5.4.5 **Corollary.** Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be any two vector norms on a finite-dimensional real or complex vector space V. Then there exist finite positive constants C_m and C_M such that $C_m\|x\|_\alpha \leq \|x\|_\beta \leq C_M\|x\|_\alpha$ for all $x \in V$.

Exercise. How does (5.4.5) break down for vector seminorms?

Exercise. Let $x=[x_1, x_2]^T \in \mathbf{R}^2$ and consider the following norms on $\mathbf{R}^2$: $\|x\|_\alpha \equiv \|[10x_1, x_2]^T\|_\infty$, and $\|x\|_\beta \equiv \|[x_1, 10x_2]^T\|_\infty$. Show that the function $f(x) \equiv (\|x\|_\alpha \|x\|_\beta)^{1/2}$ is a pre-norm on $\mathbf{R}^2$ that is not a norm. See Problem 15 at the end of this section. *Hint:* Consider $f([1,1]^T)$, $f([0,1]^T)$, and $f([1,0]^T)$.

Exercise. If $\|\cdot\|_{\alpha_1}, \ldots, \|\cdot\|_{\alpha_k}$ are vector norms on V, show that $f(x) \equiv (\|x\|_{\alpha_1} \cdots \|x\|_{\alpha_k})^{1/k}$ and $h(x) \equiv \min\{\|x\|_{\alpha_1}, \ldots, \|x\|_{\alpha_k}\}$ are pre-norms on V that are not necessarily vector norms.

One consequence of (5.4.5) is the fact that convergence (in norm) of a sequence of vectors in a finite-dimensional complex vector space is independent of the norm used.

5.4.6 **Corollary.** If $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are vector norms on a finite-dimensional real or complex vector space, and if $\{x^{(k)}\}$ is a given sequence of vectors, then $\lim_{k\to\infty} x^{(k)}=x$ with respect to $\|\cdot\|_\alpha$ if and only if $\lim_{k\to\infty} x^{(k)}=x$ with respect to $\|\cdot\|_\beta$.

Proof: Since $C_m\|x^{(k)}-x\|_\alpha \leq \|x^{(k)}-x\|_\beta \leq C_M\|x^{(k)}-x\|_\alpha$ for all k, it follows that $\|x^{(k)}-x\|_\alpha \to 0$ if and only if $\|x^{(k)}-x\|_\beta \to 0$ as $k \to \infty$. □

5.4.7 **Definition.** Two norms are said to be *equivalent* if whenever a sequence $\{x^{(k)}\}$ converges to a vector x with respect to the first norm, then it converges to the same vector with respect to the second norm. Thus, (5.4.6) says that *for finite-dimensional real or complex vector spaces, all vector norms are equivalent.* We have seen in Example (5.4.2) that two different norms might not be equivalent on an infinite-dimensional space.

Since all vector norms on $\mathbf{R}^n$ or $\mathbf{C}^n$ are equivalent to $\|\cdot\|_\infty$, we have $\lim_{k\to\infty} x^{(k)}=x$ with respect to any vector norm if and only if

$$\lim_{k\to\infty} x_i^{(k)}=x_i \quad \text{for all} \quad i=1, \ldots, n$$

Componentwise convergence (with respect to any basis) is equivalent to convergence with respect to any vector norm.

Another important consequence of equivalence of all vector norms in the finite-dimensional case is that the unit ball and unit sphere of every vector norm is compact. This fact implies that a continuous complex-valued function on the unit ball of any vector norm is bounded and that it achieves its maximum and minimum if it is real-valued.

5.4.8 **Corollary.** Let V be $\mathbf{R}^n$ or $\mathbf{C}^n$. Let $f(\cdot)$ be a pre-norm on V. The sets $\{x: f(x) \leq 1\}$ and $\{x: f(x)=1\}$ are compact. In particular, if $\|\cdot\|$ is a vector norm on V, then the closed unit ball $\{x: \|x\| \leq 1\}$ and the unit sphere $\{x: \|x\|=1\}$ are both compact.

Proof: By (5.4.4) there is some $C>0$ such that $\|x\|_2 \leq Cf(x)$ for all $x \in V$, so the set $\{x: f(x) \leq 1\}$ is a bounded set, which is contained in an ordinary Euclidean ball of radius C centered at the origin. Both of the sets $\{x: f(x)=1\}$ and $\{x: f(x) \leq 1\}$ are closed because $f(\cdot)$ is continuous. Since a closed bounded set in R^n or C^n is compact, we are done. □

Often we are not confronted with the problem of determining whether a given sequence $\{x^{(k)}\}$ converges to a given vector x, but rather with determining whether a given sequence $\{x^{(k)}\}$ converges to anything at all. For this reason, one needs to have a convergence criterion that is independent of the limit x to which the sequence converges. If there were such a limit x, then

$$\|x^{(k)}-x^{(j)}\|=\|x^{(k)}-x+x-x^{(j)}\|\le\|x^{(k)}-x\|+\|x-x^{(j)}\|\to 0$$

as $k, j\to\infty$. This is the motivation for the following.

5.4.9 **Definition.** A sequence $\{x^{(k)}\}$ in a vector space V is a *Cauchy sequence* with respect to the vector norm $\|\cdot\|$ if for each $\epsilon>0$ there is a positive integer $N(\epsilon)$ such that

$$\|x^{(k_1)}-x^{(k_2)}\|\le\epsilon$$

whenever $k_1, k_2\ge N(\epsilon)$.

5.4.10 **Theorem.** Let $\|\cdot\|$ be a given norm on a finite-dimensional real or complex vector space V, and let $\{x^{(k)}\}$ be a given sequence of vectors in V. The sequence $\{x^{(k)}\}$ converges to a vector in V if and only if it is a Cauchy sequence with respect to the norm $\|\cdot\|$.

Proof: By choosing a basis $\mathcal{B}$ of V and considering the equivalent norm $\|[x]_{\mathcal{B}}\|_\infty$, we see that there is no loss of generality if we assume that $V=\mathbf{R}^n$ or $\mathbf{C}^n$ for some integer n and if we assume that the norm is $\|\cdot\|_\infty$. If $\{x^{(k)}\}$ is a Cauchy sequence, then so is each component sequence $\{x_i^{(k)}\}$ of real or complex numbers for each $i=1,\ldots,n$. Since a Cauchy sequence of real or complex numbers must have a limit, this means that for each $i=1,\ldots,n$ there is a scalar x_i such that $\lim_{k\to\infty}x_i^{(k)}=x_i$; it is easy to check that $\lim_{k\to\infty}x^{(k)}=x$, where $x=[x_1,\ldots,x_n]^T$. On the other hand, if there is an x such that $\lim_{k\to\infty}x^{(k)}=x$, then $\|x^{(k_1)}-x^{(k_2)}\|\le\|x^{(k_1)}-x\|+\|x-x^{(k_2)}\|$ and the given sequence is a Cauchy sequence. □

It is a fundamental property of the real and complex fields (used in the proof of the preceding theorem) that a sequence is a Cauchy sequence if and only if it converges to some (real or complex) scalar. This is known as the *completeness property* of the real and complex fields, and we have just shown that the completeness property extends to finite-dimensional real and complex vector spaces with respect to any vector norm. Unfortunately, the completeness property need not hold for vector spaces that are not finite-dimensional.

5.4.11 **Definition.** A vector space V with a norm $\|\cdot\|$ is said to be *complete with respect to the norm* $\|\cdot\|$ if every sequence that is a Cauchy sequence with respect to the norm $\|\cdot\|$ converges to a point of V.

Exercise. Consider the vector space $C[0,1]$ with the L_1 norm $\|f\|_1=\int_0^1|f(t)|\,dt$, and consider the sequence of functions $\{f_k\}$ defined by

$$f_k(t)=0,\qquad 0\le t\le\frac{1}{2}-\frac{1}{k}$$

$$f_k(t)=\frac{k}{2}\left(t-\frac{1}{2}+\frac{1}{k}\right),\qquad \frac{1}{2}-\frac{1}{k}\le t\le\frac{1}{2}+\frac{1}{k}$$

$$f_k(t)=1,\qquad \frac{1}{2}+\frac{1}{k}\le t\le 1$$

Sketch the functions f_k. Show that $\{f_k\}$ is a Cauchy sequence but that there is no function $f\in C[0,1]$ for which $\lim_{k\to\infty}f_k=f$ with respect to $\|\cdot\|_1$.

Using the fact that the unit ball of any vector norm or prenorm on $\mathbf{R}^n$ or $\mathbf{C}^n$ is compact, we can introduce another useful method of generating new norms from old ones.

5.4.12 **Definition.** Let $f(\cdot)$ be a pre-norm on $V=\mathbf{R}^n$ or $\mathbf{C}^n$. The function

$$f^D(y)\equiv\max_{f(x)=1}\operatorname{Re}y^*x$$

is called the *dual norm* of f.

Observe first that the dual norm is a well-defined function on V because $\operatorname{Re}y^*x$ is a continuous function of x for each fixed $y\in V$, and the set $\{x: f(x)=1\}$ is a compact set by (5.4.8). By the Weierstrass theorem, the maximum of $\operatorname{Re}y^*x$ is attained at some point $x_0\in\{x: f(x)=1\}$. If c is a scalar such that $|c|=1$, then by the homogeneity of f we have

$$\begin{aligned}\max_{f(x)=1}|y^*x| &= \max_{f(x)=1}\max_{|c|=1}\operatorname{Re}cy^*x\\ &= \max_{f(x)=1}\max_{|c|=1}\operatorname{Re}y^*(cx)\\ &= \max_{|c|=1}\max_{f(x/c)=1}\operatorname{Re}y^*x=\max_{f(x)=1}\operatorname{Re}y^*x\end{aligned}$$

so an equivalent and sometimes convenient definition for the dual norm is

$$f^D(y)=\max_{f(x)=1}|y^*x| \tag{5.4.12a}$$

Finally, we must observe that the name *dual norm* for the function f^D is well deserved. The function $f^D(\cdot)$ is evidently homogeneous and it is positive, for if $y\neq 0$, we can use the homogeneity of $f(\cdot)$ to show that

$$f^D(y)=\max_{f(x)=1}|y^*x|\ge\left|y^*\frac{y}{f(y)}\right|=\frac{\|y\|_2^2}{f(y)}>0$$

It is perhaps remarkable that even if the function $f(\cdot)$ does not obey the triangle inequality, its dual $f^D(\cdot)$ always does:

$$f^D(y+z) = \max_{f(x)=1} |(y+z)^*x| \le \max_{f(x)=1} [|y^*x| + |z^*x|]$$
$$\le \max_{f(x)=1} |y^*x| + \max_{f(x)=1} |z^*x| = f^D(y) + f^D(z)$$

The dual norm of a pre-norm is therefore always a norm.

Thus, any pre-norm generates a norm by the process of constructing the dual norm. The most common instance of this construction is for a pre-norm that is actually a norm.

A simple inequality for the dual norm is given in the following lemma. We shall see that it is a natural generalization of the Cauchy–Schwarz inequality.

5.4.13 **Lemma.** Let $f(\cdot)$ be a pre-norm on $V = \mathbf{C}^n$ or $\mathbf{R}^n$. Then

$$|y^*x| \le f(x) f^D(y)$$
$$|y^*x| \le f^D(x) f(y)$$

for all $x, y \in V$.

Proof: If $x \ne 0$, then

$$\left| y^* \frac{x}{f(x)} \right| \le \max_{f(z)=1} |y^*z| = f^D(y)$$

and hence $|y^*x| \le f(x) f^D(y)$. Since this inequality also holds for $x = 0$, we are done. The second inequality follows from the first since $|y^*x| = |x^*y|$. □

It is easy to identify the duals of some of the most common vector norms. If $x, y \in \mathbf{C}^n$, then a special case of Hölder's inequality is

$$|y^*x| = \left| \sum_{i=1}^{n} \bar{y}_i x_i \right| \le \sum_{i=1}^{n} |\bar{y}_i x_i| \le \max_{1 \le i \le n} |y_i| \sum_{j=1}^{n} |x_j| = \|y\|_\infty \|x\|_1 \qquad (5.4.14)$$

If y is a given vector, then equality holds in (5.4.14) when x is a unit vector (with respect to $\|\cdot\|_1$) such that $x_i = 1$ for some one value of i for which $|y_i| = \|y\|_\infty$, and $x_i = 0$ otherwise. Similarly, if x is a given nonzero vector, then equality holds in (5.4.14) when y is a unit vector (with respect to $\|\cdot\|_\infty$) such that $y_i = x_i/|x_i|$ for all i such that $x_i \ne 0$ and, $y_i = 0$ otherwise. Thus,

$$(\|y\|_1)^D = \max_{\|x\|_1=1} |y^*x| = \max_{\|x\|_1=1} \|y\|_\infty \|x\|_1 = \|y\|_\infty$$
$$(\|y\|_\infty)^D = \max_{\|x\|_\infty=1} |y^*x| = \max_{\|x\|_\infty=1} \|y\|_1 \|x\|_\infty = \|y\|_1$$

We conclude that $(\|\cdot\|_1)^D = \|\cdot\|_\infty$ and $(\|\cdot\|_\infty)^D = \|\cdot\|_1$.

If we consider the Euclidean norm $\|\cdot\|_2$, a given nonzero vector y, and an arbitrary vector x, then the Cauchy–Schwarz inequality says that

$$|y^*x| = \left| \sum_{i=1}^{n} \bar{y}_i x_i \right| \le \|y\|_2 \|x\|_2 \qquad (5.4.15)$$

with equality when $x = y/\|y\|_2$. Using the same argument as above for the l_1 and l_∞ norms, we find that $(\|y\|_2)^D = \|y\|_2$, so the Euclidean norm is its own dual.

Exercise. Explain why the inequalities in (5.4.13) are a generalization of the Cauchy–Schwarz inequality (5.1.4).

Notice that for each of the three norms just considered (l_1, l_2, and l_∞), the dual of the dual norm is the original norm. This is no accident; the duality theorem (5.5.14) says that this always happens.

Among these three examples, the only norm that equals its dual is the Euclidean norm. It is not difficult to show that this is also no accident.

5.4.16 **Theorem.** Let $\|\cdot\|$ be a vector norm on $V = \mathbf{R}^n$ or $\mathbf{C}^n$, let $\|\cdot\|^D$ be its dual norm, and let $c > 0$ be given. Then $\|x\| = c\|x\|^D$ for all $x \in V$ if and only if $\|\cdot\| = \sqrt{c}\|\cdot\|_2$. In particular, $\|\cdot\| = \|\cdot\|^D$ if and only if $\|\cdot\|$ is the Euclidean norm $\|\cdot\|_2$.

Proof: If $\|\cdot\| = \sqrt{c}\|\cdot\|_2$ and $x \in V$, then

$$\|x\|^D = \max_{\|y\|=1} |x^*y| = \max_{\|y\|_2 = 1/\sqrt{c}} |x^*y| = \max_{\|y\|_2=1} \left| x^* \frac{y}{\sqrt{c}} \right|$$
$$= \frac{1}{\sqrt{c}} \max_{\|y\|_2=1} |x^*y| = \frac{1}{\sqrt{c}} \|x\|_2^D = \frac{1}{\sqrt{c}} \|x\|_2 = \frac{1}{c} \|x\|$$

for any $x \in V$. Conversely, if $\|\cdot\| = c\|\cdot\|^D$ for some $c > 0$ and if $x \in V$, then (5.4.13) gives the inequality

$$\|x\|_2^2 = |x^*x| \le \|x\| \|x\|^D = \frac{1}{c} \|x\|^2$$

so $\|x\| \ge \sqrt{c}\|x\|_2$. We can use this inequality to establish the reverse bound if $x \ne 0$ by considering

$$\frac{1}{c}\|x\| = \|x\|^D = \max_{\|y\|=1} |x^*y| = \max_{y \ne 0} \left| x^* \frac{y}{\|y\|} \right|$$
$$= \max_{y \ne 0} \left| x^* \frac{y}{\|y\|_2} \right| \frac{\|y\|_2}{\|y\|} \le \max_{y \ne 0} \left| x^* \frac{y}{\|y\|_2} \right| \frac{1}{\sqrt{c}}$$
$$= x^* \frac{x}{\|x\|_2} \frac{1}{\sqrt{c}} = \|x\|_2 \frac{1}{\sqrt{c}}$$

where we have used the fact that $\|y\|_2/\|y\| \le 1/\sqrt{c}$ for all $y \ne 0$; the Cauchy–Schwarz inequality guarantees that the maximum absolute value of the inner product between a fixed nonzero vector and a Euclidean unit vector occurs when the unit vector is parallel to the given vector. Thus, $\|x\| \le \sqrt{c}\|x\|_2$ for all $x \in V$, which, together with the reverse inequality that we have already proved, shows that $\|x\| = \sqrt{c}\|x\|_2$ for all $x \in V$. The final assertion follows when $c = 1$ and shows that the Euclidean norm is the only norm that equals its dual. □

As final remark, we observe that there is a useful sense in which a *vector*, as well as a vector norm, has a dual.

5.4.17 **Definition.** Let $x \in \mathbf{C}^n$ be a given vector and let $\|\bullet\|$ be a given vector norm on $\mathbf{C}^n$. The set

$$\{y \in \mathbf{C}^n : \|y\|^D \|x\| = y^* x = 1\}$$

is said to be *the dual of x with respect to* $\|\bullet\|$. An ordered pair of vectors $(x, y) \in \mathbf{C}^n \times \mathbf{C}^n$ is said to be a *dual pair with respect to* $\|\bullet\|$ if y is in the dual of x with respect to the norm $\|\bullet\|$.

It follows from Corollary (5.5.15) that if $\|\bullet\|$ is a vector norm, then the dual of every vector $x \in \mathbf{C}^n$ with respect to $\|\bullet\|$ is nonempty. It could consist of one point or many. If $\|\bullet\| = \|\bullet\|_2$, for example, then the dual of every vector $x \in \mathbf{C}^n$ is the one vector x itself. If $\|\bullet\| = \|\bullet\|_\infty$, on the other hand, the dual of $x = [0, 1]^T$ consists of a single vector, but the dual of $x = [1, 1]^T$ contains infinitely many vectors. See Problem 13.

Problems

1. Note that (5.4.5) may be stated equivalently as

$$C_m(\|\bullet\|_\alpha, \|\bullet\|_\beta) \le \frac{\|x\|_\beta}{\|x\|_\alpha} \le C_M(\|\bullet\|_\alpha, \|\bullet\|_\beta)$$

where $C_m(\bullet, \bullet)$ and $C_M(\bullet, \bullet)$ denote the best possible constants relating the respective norms in (5.4.5). Show that $C_m(\|\bullet\|_\beta, \|\bullet\|_\alpha) = C_M(\|\bullet\|_\alpha, \|\bullet\|_\beta)^{-1}$.

2. Express $C_m(\|\bullet\|_\alpha, \|\bullet\|_\gamma)$ in terms of $C_m(\|\bullet\|_\alpha, \|\bullet\|_\beta)$ and $C_m(\|\bullet\|_\beta, \|\bullet\|_\gamma)$, where the constants involved need not be best possible. Do likewise for C_M.

3. Verify that the accompanying table gives the best bounds $C_M(\|\bullet\|_\alpha, \|\bullet\|_\beta)$ between the l_1, l_2, and l_∞ norms; that is, $\|x\|_\alpha \le C_M \|x\|_\beta$ for all $x \in \mathbf{C}^n$ and for $\alpha, \beta = 1, 2, \infty$. In each case show that the bound is best possible by exhibiting a nonzero vector x such that $\|x\|_\alpha = C_M \|x\|_\beta$.

$\alpha \backslash \beta$	1	2	∞
1	1	$\sqrt{n}$	n
2	1	1	$\sqrt{n}$
∞	1	1	1

What is the table of best lower bounds $\|x\|_\alpha \ge C_m \|x\|_\beta$? *Hint:* See Problem 1.

4. Show that if two norms on a real or complex vector space are equivalent, then they are related by two constants and an inequality as in (5.4.5). *Hint:* Consider $f(x) = 1/\|x\|_\alpha$ on the unit sphere S of $\|\bullet\|_\beta$. If f is unbounded on S, there is a sequence $\{x_N\} \subset S$ with $\|x_N\|_\alpha < 1/N$ and $\|x_N\|_\beta \equiv 1$, which contradicts equivalence of $\|\bullet\|_\alpha$ and $\|\bullet\|_\beta$. Notice that this has nothing to do with finite dimensionality or compactness.

5. Show that the functions f_k of (5.4.2) have the property that $f(x) \to 0$ for each x, $\|f_k - f_j\|_1 \to 0$ as $k, j \to \infty$, and for each $k \ge 2$ there is some $J > k$ for which $\|f_k - f_j\|_\infty > k^{1/2}$ for all $j > J$. Thus, a sequence can be convergent in one sense (point-wise), Cauchy in a norm, and not Cauchy in another norm.

6. Let V be a complete real or complex vector space, let $\{x^{(k)}\}$ be a given sequence in V, and let $\|\bullet\|$ be a given vector norm on V. If there is an $M \ge 0$ such that $\sum_{k=1}^n \|x^{(k)}\| \le M$ for all $n = 1, 2, \ldots$, show that the sequence of partial sums $\{y^{(n)}\}$ defined by $y^{(n)} = \sum_{k=1}^n x^{(k)}$ converges to a point of V. What theorem about convergence of infinite series of real numbers does this generalize?

7. Show that $\|x\|_\infty = \lim_{p \to \infty} \|x\|_p$ for every $x \in \mathbf{C}^n$.

8. If $\alpha > 0$ and $\|\bullet\|_\alpha \equiv \alpha \|\bullet\|$, show that $(\|\bullet\|_\alpha)^D = (1/\alpha)\|\bullet\|^D$.

9. Show that the dual norm of the l_p norm is the l_q norm for any $p \ge 1$, where q is defined by the relation $1/p + 1/q = 1$. *Hint:* Replace (5.4.14) with the general form of Hölder's inequality.

10. Let $\|\bullet\|_\alpha$ and $\|\bullet\|_\beta$ be two given vector norms on $\mathbf{C}^n$, and suppose there is some $C > 0$ such that $\|x\|_\alpha \le C\|x\|_\beta$ for all $x \in \mathbf{C}^n$. Show that $\|x\|_\beta^D \le C\|x\|_\alpha^D$ for all $x \in \mathbf{C}^n$. *Hint:*

$$\|x\|_\alpha^D = \max_{\|y\|_\alpha = 1} |y^* x| = \max_{y \ne 0} \left| \frac{y^* x}{\|y\|_\alpha} \right|$$

$$= \max_{y \ne 0} \frac{|y^* x|}{\|y\|_\alpha} \ge \max_{y \ne 0} \frac{|y^* x|}{C\|y\|_\beta} = \frac{1}{C} \max_{\|y\|_\beta = 1} |y^* x|$$

11. Show that the group of isometries of $\|\cdot\|^D$ always contains the set of adjoints of all the isometries of $\|\cdot\|$. Deduce from this that the group of isometries of $\|\cdot\|^D$ is exactly the set of adjoints of the group of isometries of $\|\cdot\|$. When will the isometries of $\|\cdot\|$ and $\|\cdot\|^D$ be the same?

12. Let $\|\cdot\|$ be a vector norm on $\mathbf{C}^n$, and let $T \in M_n$. Show that $\|\cdot\|_T^D = \|\cdot\|^D$ if T is an isometry with respect to $\|\cdot\|$.

13. Let $\|\cdot\|$ be a given vector norm on $\mathbf{C}^n$. (a) Show that the dual of 0 with respect to $\|\cdot\|$ is always $\{0\}$. (b) Use Corollary (5.5.15) to show that the dual of every $x \in \mathbf{C}^n$ is nonempty. *Hint:* If $\|y_0\|^D = 1$ and $y_0^* x = \|x\|$, determine $c \geq 0$ for which $y = cy_0$ is in the dual of x. (c) Let $\|\cdot\|$ be the Euclidean norm $\|\cdot\|_2$. Show that the dual of every $x \in \mathbf{C}^n$ is $\{x\}$. (d) Let $\|\cdot\| = \|\cdot\|_\infty$. Show that the dual of $x = \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]$ is $\{x\}$; the dual of $x = \left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$ is the two line segments from $\left[\begin{smallmatrix} 0 \\ 2 \end{smallmatrix}\right]$ to $\left[\begin{smallmatrix} 2 \\ 2 \end{smallmatrix}\right]$ to $\left[\begin{smallmatrix} 2 \\ 0 \end{smallmatrix}\right]$. (e) Let $\|\cdot\| = \|\cdot\|_1$. Show that the dual of $x = \left[\begin{smallmatrix} 2 \\ 2 \end{smallmatrix}\right]$ is $\left\{\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]\right\}$ and the dual of $x = \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]$ is the two line segments from $-\left[\begin{smallmatrix} -1 \\ 1 \end{smallmatrix}\right]$ to $\left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]$ to $-\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$. (f) Show that y is in the dual of x with respect to $\|\cdot\|$ if and only if x is in the dual of y with respect to $\|\cdot\|^D$. (g) Show that the dual of x with respect to $\|\cdot\|$ is $\{x\}$ for every $x \in \mathbf{C}^n$ if and only if $\|\cdot\| = \|\cdot\|_2$.

14. Consider the function $f: \mathbf{R}^2 \to \mathbf{R}$ given by $f(x) = |x_1 x_2|^{1/2}$. Show that the set $\{x: f(x) = 1\}$ is not compact. Does this contradict (5.4.8)?

15. Consider the example of a pre-norm $f(x) = (\|x\|_\alpha \|x\|_\beta)^{1/2}$ on $\mathbf{R}^2$ given in the text, where

$$\|x\|_\alpha = \|[10x_1, x_2]^T\|_\infty, \qquad \|x\|_\beta = \|[x_1, 10x_2]^T\|_\infty$$

Show that the portion of the "unit ball" $\{x \in \mathbf{R}^n: f(x) \leq 1\}$ in the first quadrant is bounded by segments of the lines $x_2 = 1/\sqrt{10}$ and $x_1 = 1/\sqrt{10}$ and an arc of the hyperbola $x_1 x_2 = \frac{1}{100}$. Sketch this set and show that it is not convex. Why is the rest of the "unit ball" in the three remaining quadrants obtained by successive reflections of this set across the axes? Show that the unit ball of the dual norm $\{x \in \mathbf{R}^2: f^D(x) \leq 1\}$ is bounded in the first quadrant by segments of the lines $x_1/10 + x_2 = \sqrt{10}$ and $x_1 + x_2/10 = \sqrt{10}$, that the whole unit ball of f^D is obtained by successive reflections of the portion in the first quadrant, and that it is convex. Show that the portion of the unit ball of f^{DD} in the first quadrant is bounded by segments of the lines $x_2 = 1/\sqrt{10}$, $x_1 = 1/\sqrt{10}$, and $x_1 + x_2 = 11/(10\sqrt{10})$, that the rest of the unit ball is obtained by successive reflections of this set across the axes, and that it is convex. Finally, compare the unit ball of f^{DD} with that of f and show that the former is exactly the closed convex hull of the latter.

16. Let $\|\cdot\|$ be a vector norm on $V = \mathbf{R}^n$ or $\mathbf{C}^n$. Show that

$$\max_{\|x\| \neq 0} \frac{\|x\|^D}{\|x\|} = \max_{\|x\|=1} \max_{\|y\|=1} \left(\frac{x}{\|x\|_2}\right)^* \left(\frac{y}{\|y\|_2}\right) \|x\|_2 \|y\|_2$$

$$\leq \left[\max_{\|x\|=1} \|x\|_2\right]^2 \equiv C_M$$

and that

$$\min_{x \neq 0} \frac{\|x\|^D}{\|x\|} \geq \left[\min_{\|x\|=1} \|x\|_2\right]^2 \equiv C_m$$

Deduce that $C_m \|x\| \leq \|x\|^D \leq C_M \|x\|$ for all $x \in V$, so geometrical constants give bounds between every norm and its dual.

17. Let $f(\cdot)$ be a pre-norm on $\mathbf{R}^n$ or $\mathbf{C}^n$. Show that

$$f^D(y) = \max_{f(x) \leq 1} \operatorname{Re} y^* x = \max_{f(x) \leq 1} |y^* x|$$

$$= \max_{x \neq 0} \frac{\operatorname{Re} y^* x}{f(x)} = \max_{x \neq 0} \frac{|y^* x|}{f(x)} \tag{5.4.18}$$

See the exercise after (5.6.1) for another instance of this idea.

Further Readings. See [Hou 64] for a discussion of dual norms. The idea that the dual of a pre-norm is a norm seems to be due to J. Von Neumann, who discussed "gauge functions" (what we call vector norms) in "Some Matrix-Inequalities and Metrization of Matric-Space," *Tomsk Univ. Rev.* 1 (1937), 205–218. A more readily available source for this paper may be vol. 4 of Von Neumann's *Collected Works,* ed. A. H. Taub, Macmillian, New York, 1962.

5.5 Geometric properties of vector norms

The primary geometric feature of a vector norm is its unit ball, through which considerable insight about the norm may be gained.

5.5.1 **Definition.** Let $\|\cdot\|$ be a vector norm on the real or complex vector space V, let x be a point of V, and let $r > 0$ be given. The *ball of radius r around* x is the set

$$B_{\|\cdot\|}(r; x) \equiv \{y \in V: \|y - x\| \leq r\}$$

The *unit ball* of $\|\cdot\|$ is the set

$$B_{\|\cdot\|} \equiv B_{\|\cdot\|}(1; 0) = \{y \in V: \|y\| \leq 1\}$$

Exercise. Show that for every $r > 0$ and for every $x \in V$, $B(r; x) = \{y + x: y \in B(r; 0)\} = x + B(r; 0)$.

A ball of given radius around any point x looks the same as a ball of the same radius around zero; it is just translated to the point x. The unit ball is a geometric summary of a norm, which, because of the homogeneity property, characterizes the norm (actually only the boundary of $B_{\|\cdot\|}$ is needed). Here we determine exactly which subsets of $\mathbf{C}^n$ can be the unit ball of some vector norm.

Exercise. Sketch the unit balls for the l_1, l_2, and l_∞ norms on $\mathbf{R}^2$. Are there any containment relationships? Which points must be on the boundary of the unit ball of any l_p norm on $\mathbf{R}^2$? Sketch the unit ball of some other l_p norms.

Exercise. If $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are two vector norms on a vector space V, show that $\|x\|_\alpha \le \|x\|_\beta$ for all $x \in V$ if and only if $B_{\|\cdot\|_\beta} \subset B_{\|\cdot\|_\alpha}$. The natural partial order on vector norms may therefore be expressed in terms of geometric containment. What happens to the unit ball when a norm is multiplied by a positive constant?

Exercise. If $\|\cdot\|$ is a vector norm on V, if $x \in V$, and if α is a scalar such that $\|\alpha x\| = \|x\|$, show that either $x = 0$ or $|\alpha| = 1$. Conclude that each "ray" $\{\alpha x : \alpha > 0\}$ intersects the boundary of the unit ball of $\|\cdot\|$ exactly once.

5.5.2 **Definition.** A vector norm is called *polyhedral* if its unit ball is a polyhedron.

Exercise. Which of the l_p norms are polyhedral?

Exercise. If $\|\cdot\|$ is a polyhedral norm and if $S \in M_n$ is nonsingular, is $\|\cdot\|_S$ polyhedral?

The basic topological notions of open and closed sets are very easy to define in a vector space that has a norm.

5.5.3 **Definition.** Let $\|\cdot\|$ be a norm on the real or complex vector space V, and let S be a subset of V. A point $x \in S$ is said to be an *interior point* of S if there is some $\epsilon > 0$ such that $B(\epsilon; x) \subset S$. The set S is said to be *open* if every point of S is an interior point; S is said to be *closed* if its complement is open. A *limit point* of S is a point $x \in V$ such that $\lim_{k\to\infty} x^{(k)} = x$ (with respect to $\|\cdot\|$) for some sequence $\{x^{(k)}\} \subset S$. The *closure* of S is the union of S with the set of limit points of S. The *boundary* of S is the intersection of the closure of S with the closure of the complement of S. The set S is *bounded* if there exists some $M > 0$ such that $S \subset B_{\|\cdot\|}(M; 0)$. The set S is *compact* if from every covering $\bigcup_\alpha S_\alpha \supset S$ by open sets S_α one can extract finitely many sets $S_{\alpha_1}, \ldots, S_{\alpha_N}$ such that $\bigcup_{i=1}^N S_{\alpha_i} \supset S$.

Exercise. Show that the unit ball $B_{\|\cdot\|}$ is closed and bounded for any vector norm $\|\cdot\|$ on any real or complex vector space V.

Exercise. Let V be a finite-dimensional real or complex vector space, and let $S \subset V$ be a closed bounded set. Use the fact that V is isomorphic to $\mathbf{R}^n$ or $\mathbf{C}^n$ for some n (See Appendix E) to show that S is compact.

5.5.4 **Observation.** If $\|\cdot\|$ is a vector norm on a nontrivial (i.e., not zero-dimensional) real or complex vector space V, then 0 is an interior point of the unit ball $B_{\|\cdot\|}$. This follows from the homogeneity and positivity of the norm $\|\cdot\|$, which implies that $B_{\|\cdot\|}(\frac{1}{2}; 0) \subset B_{\|\cdot\|}(1; 0)$; with the boundary of the former being in the interior of the latter.

5.5.5 **Observation.** The unit ball of a vector norm is *equilibrated*; that is, if x is in the unit ball, then so is αx for all scalars α such that $|\alpha| = 1$. This follows from the homogeneity property of the vector norm.

5.5.6 **Observation.** The unit ball of a vector norm on a finite-dimensional vector space is *compact*. It is bounded because of the homogeneity property of vector norms and it is closed because the norm is always a continuous function. In the finite-dimensional case, a closed bounded set is compact, but this is not always true in the infinite-dimensional case. The property of compact sets that we shall use most frequently is the Weierstrass theorem (see Appendix E): A continuous real-valued function on a compact set is bounded and achieves both its supremum and infimum on the set. For this reason, we usually refer to the "max" or "min" of such a function.

Exercise. Consider the complex vector space l_2 of vectors $x = (x_i)$ with countably many components with the natural extension of the finite-dimensional l_2 norm

$$\|x\|_2 = \left(\sum_{k=1}^{\infty} |x_k|^2\right)^{1/2}$$

Show that $\|e_k - e_j\|_2 = \sqrt{2}$ for every pair of distinct unit basis vectors e_k and e_j, $k, j = 1, 2, \ldots$. Thus, no infinite subsequence of $\{e_k\}$ can be a Cauchy sequence, so there can be no convergent subsequence. Conclude that the unit ball of l_2 cannot be compact.

5.5.7 Observation. The unit ball of a vector norm is *convex.*

Proof: If $\|x\|\le 1$, $\|y\|\le 1$, and $\alpha\in[0,1]$, then

$$\|\alpha x+(1-\alpha)y\|\le\|\alpha x\|+\|(1-\alpha)y\|=\alpha\|x\|+(1-\alpha)\|y\|\le\alpha+(1-\alpha)\le 1$$

so that $\alpha x+(1-\alpha)y$ lies in the unit ball also. □

The foregoing necessary conditions on the unit ball of a norm are also sufficient to characterize a norm.

5.5.8 Theorem. A set B in a finite-dimensional real or complex vector space is the unit ball of a vector norm on V if and only if B is a (i) compact, (ii) convex, (iii) equilibrated set, (iv) with 0 as an interior point.

Proof: That conditions (i)–(iv) are necessary has already been observed. To see that they suffice for the definition of a norm, consider any nonzero point $x\in V$. Construct a ray segment $\{\alpha x: 0\le\alpha\le 1\}$ from the origin through x and define the "length" of x by the proportional distance along this ray from the origin to x with the length of the interval of the ray from the origin to the unique point on the boundary of the unit ball serving as one unit. More formally, define $\|x\|$ by

$$\|x\|=0 \qquad \text{if} \quad x=0$$
$$\|x\|=\min\left\{\frac{1}{t}: t>0 \text{ and } tx\in B\right\} \qquad \text{if} \quad x\ne 0$$

This function is well defined, finite, and positive for each nonzero vector x because B is compact and has 0 as an interior point. Using the equilibration assumption, it is easy to see that $\|\cdot\|$ is a homogeneous function, so it remains only to check that it satisfies the triangle inequality. If x and y are given nonzero vectors, then $x/\|x\|$ and $y/\|y\|$ are unit vectors that lie on the boundary of B. By convexity, the vector

$$z=\frac{\|x\|}{\|x\|+\|y\|}\frac{x}{\|x\|}+\frac{\|y\|}{\|x\|+\|y\|}\frac{y}{\|y\|}$$

must also lie in B. Therefore, $\|z\|\le 1$, and one easily computes that this is equivalent to $\|x+y\|\le\|x\|+\|y\|$. □

Exercise. Provide the details in the proof of (5.5.8), noting carefully where each of the four hypotheses is used.

All of the familiar l_p vector norms have the property that $\|x\|$ depends only on the *absolute value* of the entries of x. Moreover, each l_p norm is an *increasing function* of the absolute values of the entries of x. These two properties are not unrelated.

5.5.9 Definition. If $x=[x_i]\in\mathbf{F}^n$ ($\mathbf{R}^n$ or $\mathbf{C}^n$), we define $|x|\equiv[|x_i|]$. We say that $|x|\le|y|$ if $|x_i|\le|y_i|$ for all $i=1,\ldots,n$. A vector norm $\|\cdot\|$ on $\mathbf{F}^n$ is said to be

(a) *Monotone* if $|x|\le|y|$ implies $\|x\|\le\|y\|$ for all $x, y\in\mathbf{F}^n$
(b) *Absolute* if $\|x\|=\||x|\|$ for all $x\in\mathbf{F}^n$

5.5.10 Theorem. A vector norm $\|\cdot\|$ on $\mathbf{F}^n$ ($\mathbf{R}^n$ or $\mathbf{C}^n$) is monotone if and only if it is absolute.

Proof: If $\|\cdot\|$ is monotone, and if $x\in\mathbf{F}^n$, let $y\equiv|x|$. Then $|y|\le|x|$ and $|x|\le|y|$, so $\|y\|\le\|x\|$ and $\|x\|\le\|y\|$, and hence $\|\cdot\|$ is absolute. If $\|\cdot\|$ is absolute, let $x=[x_i]\in\mathbf{F}^n$ be a given vector, let k be a given integer with $1\le k\le n$, and let $\alpha\in[0,1]$. Then

$$\begin{aligned}
&\|[x_1,\ldots,x_{k-1},\alpha x_k,x_{k+1},\ldots,x_n]^T\|\\
&=\|\tfrac{1}{2}(1-\alpha)[x_1,\ldots,x_{k-1},-x_k,x_{k+1},\ldots,x_n]^T+\tfrac{1}{2}(1-\alpha)x+\alpha x\|\\
&\le\tfrac{1}{2}(1-\alpha)\|[x_1,\ldots,x_{k-1},-x_k,x_{k+1},\ldots,x_n]^T\|+\tfrac{1}{2}(1-\alpha)\|x\|+\alpha\|x\|\\
&=\tfrac{1}{2}(1-\alpha)\|x\|+\tfrac{1}{2}(1-\alpha)\|x\|+\alpha\|x\|=\|x\|
\end{aligned} \tag{5.5.11}$$

The assumption that the norm is absolute is used only in the penultimate equality. By repeating (5.5.11) for different components, one can show that an absolute norm has the property that

$$\|[\alpha_1 x_1,\ldots,\alpha_n x_n]^T\|\le\|[x_1,\ldots,x_n]^T\| \tag{5.5.12}$$

for every $x\in\mathbf{F}^n$ and all choices of $\alpha_k\in[0,1]$, $k=1,\ldots,n$. Finally, if $|x|\le|y|$, then for each $k=1,\ldots,n$ there are real numbers α_k and θ_k with $\alpha_k\in[0,1]$ such that $x_k=\alpha_k e^{i\theta_k}y_k$. Then using the absolute property we have

$$\begin{aligned}
\|x\|&=\|[\alpha_1 e^{i\theta_1}y_1,\ldots,\alpha_n e^{i\theta_n}y_n]\|=\|[\alpha_1|y_1|,\ldots,\alpha_n|y_n|]^T\|\\
&\le\|[|y_1|,\ldots,|y_n|]^T\|=\|y\|
\end{aligned}$$

so the norm must be monotone. □

The inequality (5.5.11) suggests a slightly weaker notion of monotonicity.

5.5.13 Definition. A vector norm $\|\cdot\|$ on $\mathbf{F}^n$ ($\mathbf{R}^n$ or $\mathbf{C}^n$) is said to be *weakly monotone* if

$$\|[x_1, \ldots, x_{k-1}, 0, x_{k+1}, \ldots, x_n]^T\| \le \|[x_1, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots, x_n]^T\|$$

for all $x \in \mathbf{F}^n$ and all $k = 1, \ldots, n$.

If a vector norm $\|\cdot\|$ is weakly monotone, and if $\alpha \in [0, 1]$, then

$$\begin{aligned}
&\|[x_1, \ldots, x_{k-1}, \alpha x_k, x_{k+1}, \ldots, x_n]^T\| \\
&\quad = \|(1-\alpha)[x_1, \ldots, x_{k-1}, 0, x_{k+1}, \ldots, x_n]^T + \alpha x\| \\
&\quad \le (1-\alpha)\|[x_1, \ldots, x_{k-1}, 0, x_{k+1}, \ldots, x_n]^T\| + \alpha\|x\| \\
&\quad \le (1-\alpha)\|x\| + \alpha\|x\| = \|x\|
\end{aligned}$$

so a weakly monotone norm satisfies the apparently stronger condition (5.5.12). Thus, if a point on the unit sphere of a weakly monotone norm is given, and if one of its coordinates is shrunk to zero, the entire line segment thus produced must be in the unit ball. A monotone norm is obviously weakly monotone, but not conversely, as the following exercises show.

Exercise. Show that the parallelogram with vertices at $\pm[2, 2]^T$ and $\pm[1, -1]^T$ is the unit ball of a vector norm on $\mathbf{R}^2$ that is not weakly monotone.

Exercise. Is the function $f(x) = |x_1 - x_2| + |x_2|$ a vector norm on $\mathbf{R}^2$? Is it monotone? Is it weakly monotone? Sketch its unit ball.

Exercise. Let $\|\cdot\|$ be an absolute norm on $\mathbf{R}^2$. Show that if $x = [x_1, x_2]^T$ is a point on the boundary of the unit ball, then so are the points $[\pm x_1, \pm x_2]^T$ (all four possible choices). Illustrate this geometric property with a sketch and exhibit a unit ball of a vector norm on $\mathbf{R}^2$ that is not absolute. What happens in $\mathbf{R}^n$?

Exercise. Sketch the polygon in $\mathbf{R}^2$ with vertices at $\pm[0, 1]^T$, $\pm[1, 0]^T$, and $\pm[1, 1]^T$. Explain why it is the unit ball of a vector norm on $\mathbf{R}^n$ that is weakly monotone but not monotone or absolute.

The convexity of the unit ball of a vector norm is a fact with many deep and sometimes startling implications. One of these is the following duality theorem, which we state generally in the context of pre-norms. The key ideas involved are very natural geometric ones, namely that the smallest closed convex set containing a given set S (the closed convex hull Co S; see Appendix B) is the intersection of all closed half-spaces (everything on one side of a hyperplane) containing S, and that if there is a point x, which, whenever S lies in a half-space, also lies in the same half-space, then x must belong to the closed convex hull of S. These simple

notions lead directly to the important fact that the second dual of a vector norm is identical to the original norm.

5.5.14 Theorem (duality theorem). Let $f(\cdot)$ be a pre-norm on $V = \mathbf{R}^n$ or $\mathbf{C}^n$, let f^D denote the dual norm of f and f^{DD} the dual norm of f^D, and let

$$B \equiv \{x \in V: f(x) \le 1\}, \qquad B'' \equiv \{x \in V: f^{DD}(x) \le 1\}$$

denote the "unit ball" of f and the unit ball of f^{DD}, respectively. Then

$$B \subset B'' = \text{Co } B$$

and hence $f^{DD}(x) \le f(x)$ for all $x \in V$. If f is a vector norm on V, then $B = B''$ and $f^{DD} = f$.

Proof: If $x \in V$ is a given vector, then (5.4.13) says that

$$|y^*x| \le f(x) f^D(y)$$

for any $y \in V$, and hence

$$f^{DD}(x) = \max_{f^D(y)=1} |y^*x| \le \max_{f^D(y)=1} f(x) f^D(y) = f(x)$$

Thus, $f^{DD}(x) \le f(x)$ for all $x \in V$, an inequality that is equivalent to the geometric statement $B \subset B''$.

To prove the second inclusion it is convenient to use the characterization (5.4.18) of the dual norm and to observe that the set $\{t \in V: \text{Re } t^*v \le 1\}$ is a general closed half-space that contains the origin. Using the definition of the dual norm, let $u \in B''$ be a given point and observe that

$$\begin{aligned}
u \in &\{t: \text{Re } t^*v \le 1 \text{ for every } v \text{ such that } f^D(v) \le 1\} \\
&= \{t: \text{Re } t^*v \le 1 \text{ for every } v \text{ such that } \text{Re } v^*w \le 1 \\
&\qquad\qquad \text{for every } w \text{ such that } f(w) \le 1\} \\
&= \{t: \text{Re } t^*v \le 1 \text{ for every } v \text{ such that } \text{Re } w^*v \le 1 \text{ for all } w \in B\}
\end{aligned}$$

This says that u lies in every closed half-space that has the property that it contains every point of B; that is, u lies in every closed half-space that contains B. Since the intersection of all such closed half-spaces is the closed convex hull of B, Co B, we conclude that $u \in \text{Co } B$. But the point $u \in B''$ was arbitrary, so $B'' \subset \text{Co } B$. Since Co B is the intersection of all convex sets containing B, we also have Co $B \subset B''$ and hence $B'' = \text{Co } B$.

If the pre-norm f is actually a norm, then its closed unit ball B is convex and $B = \text{Co } B = B''$. Since their unit balls are identical, the norms f and f^{DD} are the same. □

One application of the duality theorem is the following useful result. It is a special case of a finite-dimensional version of an important general result from functional analysis known as the Hahn–Banach theorem.

5.5.15 **Corollary.** Let $y \in \mathbf{C}^n$ be a given vector and let $\|\cdot\|$ be a given vector norm on $\mathbf{C}^n$. There exists a vector $y_0 \in \mathbf{C}^n$ such that

(a) $|(y_0)^* x| \le \|x\|$ for all $x \in \mathbf{C}^n$; and
(b) $(y_0)^* y = \|y\|$.

The vector y_0 is not necessarily unique, but $\|y_0\|^D = 1$ and $(y_0)^* y = \|y\|$.

Proof: We know that

$$\|y\| = (\|y\|^D)^D = \max_{\|z\|^D = 1} |y^* z|$$

by the duality theorem, and we know by compactness of the unit sphere of the vector norm $\|\cdot\|^D$ that the maximum is actually achieved for some [not necessarily unique] vector $z = y_0$ such that $\|y_0\|^D = 1$, so $\|y\| = |y^* y_0|$. By multiplying y_0 by a suitable factor of modulus 1 it is clear that the inner product $y^* y_0$ can be made positive and (b) is established. In general, we know from (5.4.13) that

$$|(y_0)^* x| \le \|y_0\|^D \|x\| = \|x\| \quad \text{for all} \quad x \in \mathbf{C}^n$$

The vector y_0 therefore satisfies (a) as well. Notice that (a) says that $\|y_0\|^D \le 1$ and (b) makes $\|y_0\|^D = 1$. □

Problems

1. Show that a set S is closed if and only if it contains all its limit points.

2. Show that every point of S is a limit point of S, so that the closure of S is just the set of limit points of S.

3. Give an example of a set that is both open and closed. Give an example of a set that is neither open nor closed.

4. Let S be a compact set in a real or complex vector space V with norm $\|\cdot\|$. Show that S is closed and bounded. If $\{x_\alpha\} \subset S$ is a given infinite sequence, show that there is a countable subsequence $\{x_{\alpha_i}\} \subset \{x_\alpha\}$ and a point $x \in S$ such that $\lim_{i\to\infty} x_{\alpha_i} = x$. Show that any closed subset of a compact set is compact.

5. What happens in (5.5.4) if V is zero-dimensional?

6. How might the unit ball of a vector seminorm be defined, and how does its shape differ from the unit ball of a norm? Sketch an example.

7. If $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are vector norms on a vector space and if $\|\cdot\|$ is the vector norm defined by

$$\|x\| = \max\{\|x\|_\alpha, \|x\|_\beta\}$$

show that $B_{\|\cdot\|} = B_{\|\cdot\|_\alpha} \cap B_{\|\cdot\|_\beta}$.

8. Show that a vector norm $\|\cdot\|$ on $\mathbf{F}^n$ ($\mathbf{R}^n$ or $\mathbf{C}^n$) is absolute if and only if

$$\|[\alpha_1 x_1, \alpha_2 x_2, \ldots, \alpha_n x_n]^T\| = \|[x_1, x_2, \ldots, x_n]^T\|$$

for all $[x_1, x_2, \ldots, x_n]^T \in \mathbf{F}^n$ and all scalars $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbf{F}$ such that $|\alpha_1| = \cdots = |\alpha_n| = 1$.

In the next six problems, the following notation is used. Let $x, y \in V$ and let $\|\cdot\|$ be a vector norm on the real or complex vector space V. Then

$$L(x, y) \equiv \{z(t) = x + t(y - x) : 0 \le t \le 1\}$$

denotes the usual (linear algebraic) *line segment between x and y*, and

$$C(x, y; \|\cdot\|) \equiv \{z \in V: \|x - z\| + \|z - y\| = \|x - y\|\}$$

denotes the (metric) *convex hull of x and y with respect to the norm* $\|\cdot\|$.

9. Show that $L(x, y) \subset C(x, y; \|\cdot\|)$ for all $x, y \in V$ and for any vector norm $\|\cdot\|$.

10. If $V = \mathbf{C}^n$ and if the norm is the l_2 norm, show that $C(x, y; \|\cdot\|_2) = L(x, y)$ for all $x, y \in \mathbf{C}^n$; that is, show that $\|x + y\|_2 = \|x - z\|_2 + \|z - y\|_2$ if and only if $z = x + t(y - x)$ for some $t \in [0, 1]$.

11. Show that $C(x, y; \|\cdot\|)$ is always an ordinary convex set; that is, show that if $z_1, z_2 \in C(x, y; \|\cdot\|)$, then $tz_1 + (1 - t)z_2 \in C(x, y; \|\cdot\|)$ for all $t \in [0, 1]$.

12. Consider $V = \mathbf{R}^2$ over $\mathbf{R}$ and show that $C((1, 0), (0, 1); \|\cdot\|_1)$ is the entire square whose vertices are at the points $(0, 0)$, $(0, 1)$, $(1, 1)$, and $(1, 0)$ in the plane. *Hint:* Show that $(0, 0)$ and $(1, 1)$ are in the l_1 convex hull of $(0, 1)$ and $(1, 0)$ and use Problem 11. Show that $C((1, 0), (0, 1); \|\cdot\|_\infty)$ is just the line segment $L((1, 0), (0, 1))$, however.

13. Consider $V = \mathbf{R}^2$ over $\mathbf{R}$ again and show that $C((1, 1), (1, -1); \|\cdot\|_\infty)$ is the entire square whose vertices are at the points $(0, 0)$, $(1, 1)$, $(2, 0)$, and $(1, -1)$ in the plane. *Hint:* Show that $(0, 0)$ and $(2, 0)$ are in the l_∞ convex hull of $(1, 1)$ and $(1, -1)$. Show that $C((1, 1), (1, -1); \|\cdot\|_1)$ is just the line segment $L((1, 1), (1, -1))$, however.

14. The metric convex hull of a set of k points $S \subset V$, $k \ge 2$, may be defined to be the set of all $z \in V$ such that z is in the metric convex hull of

two points, each of which is in the metric convex hull of some pair of points of S. Show that this agrees with the above definition when $k=2$ and describe the l_1 convex hull of the set of unit orthonormal basis vectors $\{e_1, e_2, \ldots, e_n\}$ in $\mathbf{R}^n$. What is the l_2 convex hull of this set? What is the ordinary linear algebraic convex hull of this set?

Further Readings. See [Hou 64] for more discussion of geometrical aspects of vector norms. The key idea for the proof of the duality theorem (the identification of the unit ball of the second dual of a norm or pre-norm with the intersection of all the half-spaces containing the unit ball of the norm or pre-norm) is used by Von Neumann in the paper cited at the end of Section (5.4). See [Val] for a detailed discussion of convex sets, convex hulls, half-spaces, and so forth.

5.6 Matrix norms

Since M_n is itself a vector space of dimension n^2, one can measure the "size" of a matrix by using any vector norm on $\mathbf{C}^{n^2}$. However, M_n is not just a high-dimensional vector space; it has a natural multiplication operation, and it is often useful in making estimates to relate the "size" of AB to the "sizes" of A and B.

We call a function $\|\cdot\|: M_n \to \mathbf{R}$ a *matrix norm* if for all $A, B \in M_n$ it satisfies the following five axioms:

(1)	$\|A\| \ge 0$	Nonnegative
(1a)	$\|A\| = 0$ if and only if $A = 0$	Positive
(2)	$\|cA\| = \lvert c\rvert\, \|A\|$ for all complex scalars c	Homogeneous
(3)	$\|A+B\| \le \|A\| + \|B\|$	Triangle inequality
(4)	$\|AB\| \le \|A\|\,\|B\|$	Submultiplicative

Notice that properties (1)–(3) are identical to the axioms for a vector norm (5.1.1). A vector norm on matrices, that is, a function that satisfies (1)–(3) and not necessarily (4), is often called a *generalized matrix norm.* The notions of a matrix seminorm and a generalized matrix seminorm may also be defined via omission of axiom (1a).

Since $\|A^2\| = \|AA\| \le \|A\|\,\|A\| = \|A\|^2$ for any matrix norm, it must be that $\|A\| \ge 1$ for any nonzero matrix A for which $A^2 = A$. In particular, $\|I\| \ge 1$ for any matrix norm. If A is invertible, then $I = AA^{-1}$, so $\|I\| = \|AA^{-1}\| \le \|A\|\,\|A^{-1}\|$, and we have the lower bound

$$\|A^{-1}\| \ge \frac{\|I\|}{\|A\|}$$

for any matrix norm $\|\cdot\|$.

Exercise. Show that if $\|\cdot\|$ is a matrix norm, then $\|A^k\| \ge \|A\|^k$ for every $k = 1, 2, \ldots$, and all $A \in M_n$. Give an example of a vector norm on matrices for which this inequality is not true.

Some of the vector norms introduced in (5.2) are matrix norms when applied to the vector space M_n and some are not. The most familiar examples are the l_p norms for $p = 1, 2, \infty$. They are already known to be vector norms, so one needs to verify only axiom (4).

Example. The l_1 *norm* defined for $A \in M_n$ by

$$\|A\|_1 \equiv \sum_{i,j=1}^{n} |a_{ij}|$$

is a matrix norm because

$$\begin{aligned}\|AB\|_1 &= \sum_{i,j=1}^{n} \left| \sum_{k=1}^{n} a_{ik} b_{kj} \right| \le \sum_{i,j,k=1}^{n} |a_{ik} b_{kj}| \\ &\le \sum_{i,j,k,m=1}^{n} |a_{ik} b_{mj}| = \left(\sum_{i,k=1}^{n} |a_{ik}| \right) \left(\sum_{j,m=1}^{n} |b_{mj}| \right) \\ &= \|A\|_1 \|B\|_1\end{aligned}$$

The first inequality comes from the triangle inequality, while the second comes from adding additional terms to the sum.

Example. The *Euclidean norm* or l_2 *norm* defined for $A \in M_n$ by

$$\|A\|_2 \equiv \left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2}$$

is a matrix norm because

$$\begin{aligned}\|AB\|_2^2 &= \sum_{i,j=1}^{n} \left| \sum_{k=1}^{n} a_{ik} b_{kj} \right|^2 \le \sum_{i,j=1}^{n} \left(\sum_{k=1}^{n} |a_{ik}|^2 \right) \left(\sum_{m=1}^{n} |b_{mj}|^2 \right) \\ &= \left(\sum_{i,k=1}^{n} |a_{ik}|^2 \right) \left(\sum_{m,j=1}^{n} |b_{mj}|^2 \right) = \|A\|_2^2 \|B\|_2^2\end{aligned}$$

This inequality is just the Cauchy–Schwarz inequality. When applied to matrices, this norm is sometimes called the *Frobenius norm,* the *Schur norm,* or the *Hilbert–Schmidt norm.* Notice that if $A = [a_1\, a_2 \cdots a_n] \in M_n$ is written in terms of its column vectors $a_i \in \mathbf{C}^n$, then

$$\|A\|_2^2 = \|a_1\|_2^2 + \cdots + \|a_n\|_2^2$$

Since the l_2 norm on $\mathbf{C}^n$ is unitarily invariant, we have the important fact that

$$\|UA\|_2^2 = \|Ua_1\|_2^2 + \cdots + \|Ua_n\|_2^2 = \|a_1\|_2^2 + \cdots + \|a_n\|_2^2 = \|A\|_2^2$$

whenever $U \in M_n$ is unitary. Since $\|B^*\|_2 = \|B\|_2$ for all $B \in M_n$, this implies that

$$\|UAV\|_2 = \|AV\|_2 = \|V^*A^*\|_2 = \|A^*\|_2 = \|A\|_2$$

whenever $U, V \in M_n$ are unitary. Thus, the l_2 norm on M_n is a unitarily invariant matrix norm.

Example. The l_∞ *norm* defined for $A \in M_n$ by

$$\|A\|_\infty \equiv \max_{1 \le i,j \le n} |a_{ij}|$$

is a norm on the vector space M_n but is not a matrix norm. Consider the matrix $J = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \in M_2$ and compute $J^2 = 2J$, $\|J\|_\infty = 1$, $\|J^2\|_\infty = \|2J\|_\infty = 2\|J\|_\infty = 2$. It is not the case that $\|J^2\|_\infty \le \|J\|_\infty^2$, and hence $\|\cdot\|_\infty$ is not a submultiplicative norm. However, if we define

$$\|\!|A|\!\| \equiv n\|A\|_\infty, \qquad A \in M_n$$

then we have

$$\|\!|AB|\!\| = n \max_{1 \le i,j \le n} \left| \sum_{k=1}^{n} a_{ik} b_{kj} \right| \le n \max_{1 \le i,j \le n} \sum_{k=1}^{n} |a_{ik} b_{kj}|$$

$$\le n \max_{1 \le i,j \le n} \sum_{k=1}^{n} \|A\|_\infty \|B\|_\infty = n\|A\|_\infty n\|B\|_\infty$$

$$= \|\!|A|\!\| \, \|\!|B|\!\|$$

Thus, only a minor modification of the vector norm $\|\cdot\|_\infty$ is required to make it a matrix norm.

Associated with each vector norm $\|\cdot\|$ on $\mathbf{C}^n$ is a natural matrix norm $\|\!|\cdot|\!\|$ that is "induced" by $\|\cdot\|$ on M_n. The norm $\|\!|\cdot|\!\|$ is constructed from $\|\cdot\|$, and this construction adds to the list of methods for producing one norm from another.

5.6.1 **Definition.** Let $\|\cdot\|$ be a vector norm on $\mathbf{C}^n$. Define $\|\!|\cdot|\!\|$ on M_n by

$$\|\!|A|\!\| \equiv \max_{\|x\|=1} \|Ax\|$$

The "max" in the above definition (rather than "sup") is justified since $\|Ax\|$ is a continuous function of x and the unit ball $B_{\|\cdot\|}$ is a compact set (see Appendix E).

Exercise. Show that the norm (5.6.1) may also be computed in the following equivalent ways:

$$\|\!|A|\!\| = \max_{\|x\|=1} \|Ax\|$$

$$= \max_{\|x\| \le 1} \|Ax\|$$

$$= \max_{x \ne 0} \frac{\|Ax\|}{\|x\|}$$

$$= \max_{\|x\|_\alpha = 1} \frac{\|Ax\|}{\|x\|}, \qquad \text{where } \|\cdot\|_\alpha \text{ is any vector norm}$$

5.6.2 **Theorem.** The function $\|\!|\cdot|\!\|$ defined in (5.6.1) is a matrix norm on M_n, $\|Ax\| \le \|\!|A|\!\| \|x\|$ for all $A \in M_n$ and all $x \in \mathbf{C}^n$, and $\|\!|I|\!\| = 1$.

Proof: Axiom (1) at the beginning of this section follows from the fact that $\|\!|A|\!\|$ is the maximum of a nonnegative valued function, and (1a) follows from the fact that $Ax = 0$ for all x precisely when $A = 0$. Axiom (2) follows from the calculation

$$\|\!|cA|\!\| = \max \|cAx\| = \max |c| \, \|Ax\| = |c| \max \|Ax\| = |c| \, \|\!|A|\!\|$$

Similarly, the triangle inequality (3) is inherited, since

$$\|\!|A+B|\!\| = \max \|(A+B)x\| = \max \|Ax + Bx\| \le \max(\|Ax\| + \|Bx\|)$$

$$\le \max \|Ax\| + \max \|Bx\| = \|\!|A|\!\| + \|\!|B|\!\|$$

The submultiplicative axiom (4) follows from the fact that

$$\|\!|AB|\!\| = \max \frac{\|ABx\|}{\|x\|} = \max \frac{\|ABx\|}{\|Bx\|} \frac{\|Bx\|}{\|x\|}$$

$$\le \max \frac{\|Ay\|}{\|y\|} \max \frac{\|Bx\|}{\|x\|} = \|\!|A|\!\| \, \|\!|B|\!\|$$

where we assume, without loss of generality, that the maximum is taken over only those x that are not in the null space of B. For the next assertion, we observe that if $x \ne 0$, then $\|Ax/\|x\|\| \le \|\!|A|\!\|$ because of the definition of this norm as a maximum. By homogeneity of the vector norm we obtain $\|Ax\| \le \|\!|A|\!\| \|x\|$, which also holds when $x = 0$. Finally,

$$\|\!|I|\!\| = \max_{\|x\|=1} \|Ix\| = \max_{\|x\|=1} \|x\| = 1$$

□

5.6.3 **Definition.** We say that the matrix norm $|||\cdot|||$ defined in (5.6.1) is the matrix norm *induced* by the vector norm $\|\cdot\|$. It is sometimes called the *operator norm* or *lub* (least upper bound) norm associated with the vector norm $\|\cdot\|$.

Notice that the operator norm is a matrix norm as a consequence of general properties of all vector norms. Therefore, one way to prove that a certain function on M_n is a matrix norm is to show that it is induced by some vector norm. We shall adopt this strategy when we discuss an important matrix norm called the spectral norm.

The inequality in the statement of Theorem (5.6.2) says that the vector norm $\|\cdot\|$ is *compatible* with the induced matrix norm $|||\cdot|||$, and this theorem shows that associated with any vector norm on $\mathbf{C}^n$ there is a compatible matrix norm on M_n. The theorem also gives the necessary condition $|||I|||=1$ for a matrix norm $|||\cdot|||$ to be induced by some vector norm; unfortunately, this necessary condition is not also sufficient.

We next note several important examples of matrix norms that are induced by familiar l_p norms but can also be calculated independent of the definition (5.6.1). In each case, we take $A=[a_{ij}]\in M_n$.

5.6.4 The *maximum column sum matrix norm* $|||\cdot|||_1$ is defined on M_n by

$$|||A|||_1 \equiv \max_{1\le j\le n} \sum_{i=1}^{n} |a_{ij}|$$

The norm $|||\cdot|||_1$ is induced by the l_1 vector norm and hence must be a matrix norm. One can show this as follows. Write $A\in M_n$ in terms of its columns as $A=[a_1 \cdots a_n]$. Then $|||A|||_1=\max_{1\le i\le n}\|a_i\|_1$. If $x=[x_i]$, then

$$\|Ax\|_1=\|x_1a_1+\cdots+x_na_n\|_1\le \sum_{i=1}^{n}\|x_ia_i\|_1=\sum_{i=1}^{n}|x_i|\,\|a_i\|_1$$

$$\le \sum_{i=1}^{n}|x_i|\left(\max_{1\le k\le n}\|a_k\|_1\right)=\sum_{i=1}^{n}|x_i|\,|||A|||_1$$

$$=\|x\|_1\,|||A|||_1$$

Thus, $\max_{\|x\|_1=1}\|Ax\|_1\le |||A|||_1$. If we now choose $x=e_k$ (the kth unit basis vector), then for any $k=1,2,\ldots,n$ we have

$$\max_{\|x\|_1=1}\|Ax\|_1\ge \|1\,a_k\|_1=\|a_k\|_1$$

and hence

$$\max_{\|x\|_1=1}\|Ax\|_1\ge \max_{1\le k\le n}\|a_k\|_1=|||A|||_1$$

Since we have now proved that the matrix norm induced by the l_1 vector norm is both an upper bound and a lower bound on $|||A|||_1$, we are done.

Exercise. Prove directly from the definition that $|||\cdot|||_1$ is a matrix norm.

5.6.5 The *maximum row sum matrix norm* $|||\cdot|||_\infty$ is defined on M_n by

$$|||A|||_\infty \equiv \max_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}|$$

The norm $|||\cdot|||_\infty$ is induced by the l_∞ vector norm and hence must be a matrix norm. The argument is similar to the proof for the maximum column sum norm. We compute

$$\|Ax\|_\infty=\max_{1\le i\le n}\left|\sum_{j=1}^{n}a_{ij}x_j\right|\le \max_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}x_j|\le \max_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}|\,\|x\|_\infty$$

$$=|||A|||_\infty\|x\|_\infty$$

and hence $\max_{\|x\|_\infty=1}\|Ax\|_\infty\le |||A|||_\infty$. If $A=0$ there is nothing to prove, so we may assume that $A\ne 0$. Suppose the kth row of A is nonzero and define the vector $z=[z_i]\in\mathbf{C}^n$ by

$$z_i=\frac{\bar{a}_{ki}}{|a_{ki}|}\quad\text{if}\quad a_{ki}\ne 0$$

$$z_i=1\quad\text{if}\quad a_{ki}=0$$

Then $\|z\|_\infty=1$, $a_{kj}z_j=|a_{kj}|$ for all $j=1,2,\ldots,n$, and

$$\max_{\|x\|_\infty=1}\|Ax\|_\infty\ge\|Az\|_\infty=\max_{1\le i\le n}\left|\sum_{j=1}^{n}a_{ij}z_j\right|\ge\left|\sum_{j=1}^{n}a_{kj}z_j\right|=\sum_{j=1}^{n}|a_{kj}|$$

Thus,

$$\max_{\|x\|_\infty=1}\|Ax\|_\infty\ge\max_{1\le k\le n}\sum_{j=1}^{n}|a_{kj}|=|||A|||_\infty$$

and we are done.

Exercise. Verify directly from the definition that $|||\cdot|||_\infty$ is a matrix norm on M_n.

5.6.6 The *spectral norm* $|||\cdot|||_2$ is defined on M_n by

$$|||A|||_2\equiv\max\{\sqrt{\lambda}:\lambda \text{ is an eigenvalue of } A^*A\}$$

Notice that if $A^*Ax=\lambda x$ and $x\ne 0$, then $x^*A^*Ax=\|Ax\|_2^2=\lambda\|x\|_2^2$, so $\lambda\ge 0$ and $\sqrt{\lambda}$ is real and nonnegative.

Exercise. If B is a normal matrix and $B=U^*\Lambda U$ with U unitary and $\Lambda=\text{diag}(\lambda_1,\ldots,\lambda_n)$, show that

$$|x^*Bx|\leq \max\{|\lambda|:\lambda \text{ is an eigenvalue of } B\}\|x\|_2^2$$

Exercise. Show that $\|Ax\|_2^2=x^*A^*Ax$ for all $x\in\mathbf{C}^n$ and use the previous exercise to show that $|||\cdot|||_2$ is the matrix norm induced by the Euclidean vector norm $\|\cdot\|_2$. Conclude from this that the spectral norm is in fact a matrix norm.

Exercise. Show that $|||UAV|||_2=|||A|||_2$ for any $A\in M_n$ and any unitary matrices $U, V\in M_n$. Thus, the spectral norm is a unitarily invariant matrix norm.

We next show that one matrix norm may be transformed into another by a fixed similarity.

5.6.7 **Theorem.** If $|||\cdot|||$ is a matrix norm on M_n and if $S\in M_n$ is non-singular, then

$$|||A|||_S\equiv|||S^{-1}AS||| \quad \text{for all} \quad A\in M_n$$

is a matrix norm.

Proof: The axioms (1), (1a), (2), and (3) are verified in a straightforward manner for $|||\cdot|||_S$. The submultiplicativity of $|||\cdot|||_S$ follows from the calculation

$$|||AB|||_S=|||S^{-1}ABS|||=|||(S^{-1}AS)(S^{-1}BS)|||\leq|||S^{-1}AS|||\,|||S^{-1}BS|||$$
$$=|||A|||_S|||B|||_S$$

□

Theorem (5.6.7) can be of great use in tailoring a matrix norm for a specific purpose. Some applications of this type are developed here and in the following section.

One important area of application of matrix norms is in giving bounds for the spectrum of a matrix.

5.6.8 **Definition.** The *spectral radius* $\rho(A)$ of a matrix $A\in M_n$ is

$$\rho(A)\equiv\max\{|\lambda|:\lambda \text{ is an eigenvalue of } A\}$$

Observe that if λ is any eigenvalue of A, then $|\lambda|\leq\rho(A)$; moreover, there is at least one eigenvalue λ for which $|\lambda|=\rho(A)$. If $Ax=\lambda x$, $x\neq 0$, and if $|\lambda|=\rho(A)$, consider the matrix $X\in M_n$ all the columns of which are equal to the eigenvector x, and observe that $AX=\lambda X$. If $|||\cdot|||$ is any matrix norm,

$$|\lambda|\,|||X|||=|||\lambda X|||=|||AX|||\leq|||A|||\,|||X|||$$

and therefore $|\lambda|=\rho(A)\leq|||A|||$. This is a proof of the following theorem.

5.6.9 **Theorem.** If $|||\cdot|||$ is any matrix norm and if $A\in M_n$, then $\rho(A)\leq |||A|||$.

Exercise. Give an example of a vector norm $\|\cdot\|$ on matrices and a matrix $A\in M_n$ such that $\|A\|<\rho(A)$.

Exercise. Let $|||\cdot|||$ be a matrix norm on M_n, and consider the mapping $F:\mathbf{C}^n\to M_n$ defined by $F(x)=[x\;x\ldots x]=$ the matrix in M_n all of whose columns are just x. Show that the function $\|\cdot\|$ defined on $\mathbf{C}^n$ by $\|x\|\equiv|||F(x)|||$ is a norm on $\mathbf{C}^n$ and that $\|Ax\|\leq|||A|||\,\|x\|$ for all $x\in\mathbf{C}^n$ and all $A\in M_n$. This inequality says that the vector norm $\|\cdot\|$ is *compatible* with the matrix norm $|||\cdot|||$, and this exercise shows that any matrix norm on M_n has a compatible vector norm on $\mathbf{C}^n$.

Although the spectral radius function is not itself a matrix or vector norm on M_n (see Problem 19), for each fixed $A\in M_n$, it is the greatest lower bound for the values of all matrix norms of A.

5.6.10 **Lemma.** Let $A\in M_n$ and $\epsilon>0$ be given. There is a matrix norm $|||\cdot|||$ such that $\rho(A)\leq|||A|||\leq\rho(A)+\epsilon$.

Proof: By the Schur triangularization theorem (2.3.1), there is a unitary matrix U and an upper triangular matrix Δ such that $A=U^*\Delta U$. Set $D_t\equiv\text{diag}(t,t^2,t^3,\ldots,t^n)$ and compute

$$D_t\Delta D_t^{-1}=\begin{bmatrix}\lambda_1 & t^{-1}d_{12} & t^{-2}d_{13} & \ldots & t^{-n+1}d_{1n}\\ 0 & \lambda_2 & t^{-1}d_{23} & \ldots & t^{-n+2}d_{2n}\\ 0 & 0 & \lambda_3 & \ldots & t^{-n+3}d_{3n}\\ \cdot & \cdot & \cdot & \ldots & \cdot\\ 0 & 0 & 0 & \ldots & t^{-1}d_{n-1,n}\\ 0 & 0 & 0 & 0 & \lambda_n\end{bmatrix}$$

Thus, for $t>0$ large enough, we can be certain that the sum of all the absolute values of the off-diagonal entries of $D_t\Delta D_t^{-1}$ is less than ϵ. In particular, we can be sure that $|||D_t\Delta D_t^{-1}|||_1\leq\rho(A)+\epsilon$ for large enough t. Thus, if we define the matrix norm $|||\cdot|||$ by

$$|||B|||\equiv|||D_tU^*BUD_t^{-1}|||_1=|||(UD_t^{-1})^{-1}B(UD_t^{-1})|||_1$$

for any $B \in M_n$, and if we choose t large enough, then we will have constructed a matrix norm such that $\|A\| \le \rho(A)+\epsilon$. Since $\|A\| \ge \rho(A)$ for any matrix norm, we are done. □

Exercise. Explain why the preceding results show that $\rho(A) = \inf\{\|A\| : \|\cdot\| \text{ is a matrix norm}\}$.

We are interested in characterizing matrices A such that $A^k \to 0$ as $k \to \infty$. The following result is the last tool we need to attack this problem.

5.6.11 **Lemma.** Let $A \in M_n$ be a given matrix. If there is a matrix norm $\|\cdot\|$ such that $\|A\| < 1$, then $\lim_{k\to\infty} A^k = 0$; that is, all the entries of A^k tend to zero as $k \to \infty$.

Proof: If $\|A\| < 1$, then $\|A^k\| \le \|A\|^k \to 0$ as $k \to \infty$. This says that $A^k \to 0$ with respect to the norm $\|\cdot\|$, but since all vector norms on the n^2 dimensional space M_n are equivalent, it must also be the case that $A^k \to 0$ with respect to the vector norm $\|\cdot\|_\infty$. □

Exercise. Give an example of a matrix A and two matrix norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ such that $\|A\|_\alpha < 1$ and $\|A\|_\beta > 1$. Conclusion?

Matrices $A \in M_n$ such that $\lim_{k\to\infty} A^k = 0$ are called *convergent* and are important in many applications, for example, in the analysis of iterative processes. In is therefore important to be able to characterize convergent matrices.

5.6.12 **Theorem.** Let $A \in M_n$. Then $\lim_{k\to\infty} A^k = 0$ if and only if $\rho(A) < 1$.

Proof: If $A^k \to 0$ and if $x \neq 0$ is a vector such that $Ax = \lambda x$, then $A^k x = \lambda^k x \to 0$ only if $|\lambda| < 1$. Since this inequality must hold for every eigenvalue of A, we conclude that $\rho(A) < 1$. Conversely, if $\rho(A) < 1$, then by Lemma (5.6.10) there is some matrix norm $\|\cdot\|$ such that $\|A\| < 1$. Thus, $A^k \to 0$ as $k \to \infty$ by Lemma (5.6.11). □

Exercise. Consider the matrix $A = \begin{bmatrix} 1/2 & 1 \\ 0 & 1/2 \end{bmatrix} \in M_2$. Compute A^k and $\rho(A^k)$ explicitly for $k = 2, 3, \ldots$. Show that $\rho(A^k) = [\rho(A)]^k$. How do the following behave as $k \to \infty$? The entries of A^k; $\|A^k\|_1$; $\|A^k\|_\infty$; $\|A^k\|_2$.

Exercise. Let $A = \begin{bmatrix} 1/2 & 1 \\ -1/8 & 1/2 \end{bmatrix}$, and define a sequence of vectors $\{x^{(k)}\} \in \mathbf{C}^2$ by the recursion $x^{(k+1)} = Ax^{(k)}$, $k = 0, 1, \ldots$. Show that, regardless of the initial vector $x^{(0)}$ chosen, $x^{(k)} \to 0$ as $k \to \infty$.

Sometimes one needs bounds on the size of the entries of A^k as $k \to \infty$. One useful bound is an immediate consequence of the previous theorem.

5.6.13 **Corollary.** Let $A \in M_n$ be a given matrix, and let $\epsilon > 0$ be given. There is a constant $C = C(A, \epsilon)$ such that

$$|(A^k)_{ij}| \le C(\rho(A)+\epsilon)^k$$

for all $k = 1, 2, 3, \ldots$ and all $i, j = 1, 2, 3, \ldots, n$.

Proof: Since the matrix $\tilde{A} \equiv [\rho(A)+\epsilon]^{-1}A$ has spectral radius strictly less than 1, it is convergent and hence $\tilde{A}^k \to 0$ as $k \to \infty$. In particular, the elements of the sequence $\{\tilde{A}^k\}$ are bounded, so there is some finite $C > 0$ such that $|(\tilde{A}^k)_{ij}| \le C$ for all $k = 1, 2, 3, \ldots$ and all $i, j = 1, 2, \ldots, n$. This is the asserted bound. □

Exercise. Let $A = \begin{bmatrix} a & 1 \\ 0 & a \end{bmatrix}$, compute A^k explicitly, and show that one may not always take $\epsilon = 0$ in (5.6.13).

Even though it is not accurate to say that individual entries of A^k behave like $\rho(A)^k$ as $k \to \infty$, the sequence $\{\|A^k\|\}$ does have this asymptotic behavior for any matrix norm $\|\cdot\|$.

5.6.14 **Corollary.** Let $\|\cdot\|$ be a matrix norm on M_n. Then

$$\rho(A) = \lim_{k\to\infty} \|A^k\|^{1/k}$$

for all $A \in M_n$.

Proof: Since $\rho(A)^k = \rho(A^k) \le \|A^k\|$, we have that $\rho(A) \le \|A^k\|^{1/k}$ for all $k = 1, 2, \ldots$. If $\epsilon > 0$ is given, the matrix $\tilde{A} \equiv [\rho(A)+\epsilon]^{-1}A$ has spectral radius strictly less than 1 and hence it is convergent. Thus, $\|\tilde{A}^k\| \to 0$ as $k \to \infty$ and there is some $N = N(\epsilon, A)$ such that $\|\tilde{A}^k\| < 1$ for all $k \ge N$. This is just the statement that $\|A^k\| \le [\rho(A)+\epsilon]^k$ for all $k \ge N$, or that $\|A^k\|^{1/k} \le \rho(A)+\epsilon$ for all $k \ge N$. Since $\rho(A) \le \|A^k\|^{1/k}$ for all k and since $\epsilon > 0$ is arbitrary, we conclude that $\lim_{k\to\infty}\|A^k\|^{1/k}$ exists and equals $\rho(A)$. □

Questions about the convergence of infinite sequences or series of matrices can be treated with vector norms just as one treats infinite sequences or series of vectors.

Exercise. Let $\{A_k\} \subset M_n$ be a given infinite sequence of matrices. Show that the series $\sum_{k=0}^{\infty} A_k$ converges to some matrix in M_n if there is a vector norm $\|\cdot\|$ on M_n such that the numerical series $\sum_{k=0}^{\infty} \|A_k\|$ is convergent (or even if its partial sums are bounded). *Hint:* Show that the partial sums form a Cauchy sequence.

One special case for matrices that does not arise in the study of infinite series of vectors is the case of power series of matrices. But because of the submultiplicative property of matrix norms, it is easy to give a simple sufficient condition for convergence of matrix power series.

5.6.15 **Theorem.** If $A \in M_n$, then the series $\sum_{k=0}^{\infty} a_k A^k$ converges if there is a matrix norm $\|\cdot\|$ on M_n such that the numerical series $\sum_{k=0}^{\infty} |a_k| \, \|A\|^k$ converges, or even if the partial sums of this series are bounded.

Exercise. Prove (5.6.15).

Exercise. Show by example that it is possible that the series $\sum_{k=0}^{\infty} a_k A^k$ converges and the series $\sum_{k=0}^{\infty} |a_k| \, \|A\|^k$ diverges. This is analogous to conditional convergence (convergence but not absolute convergence) for numerical series.

Exercise. Let the function $f(z)$ be defined by the power series $f(z) = \sum_{k=0}^{\infty} a_k z^k$, which has radius of convergence $R > 0$, and let $\|\cdot\|$ be a matrix norm on M_n. Show that $f(A) \equiv \sum_{k=0}^{\infty} a_k A^k$ is well defined for all $A \in M_n$ such that $\|A\| < R$. More generally, show that $f(A)$ is well defined for all $A \in M_n$ such that $\rho(A) < R$.

Exercise. If A is diagonalizable and $A = S^{-1}\Lambda S$, one sometimes defines $f(A) \equiv S^{-1} f(\Lambda) S$, where $f(\Lambda) \equiv \mathrm{diag}(f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_n))$. Show that this definition of $f(A)$ agrees with the power series definition in the preceding exercise if A is diagonalizable. Is one of the two definitions more general than the other?

Exercise. Show that the matrix exponential given by the power series

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k$$

is well defined for every $A \in M_n$.

Exercise. How would you define $\cos(A)$? For what A is this defined?

5.6.16 **Corollary.** A matrix $A \in M_n$ is invertible if there is a matrix norm $\|\cdot\|$ such that $\|I - A\| < 1$. If this condition is satisfied,

$$A^{-1} = \sum_{k=0}^{\infty} (I - A)^k$$

Proof: If $\|I - A\| < 1$, then the series

$$\sum_{k=0}^{\infty} (I - A)^k$$

converges to some matrix C because the radius of convergence of the series $\sum z^k$ is 1. But since

$$A \sum_{k=0}^{N} (I - A)^k = [I - (I - A)] \sum_{k=0}^{N} (I - A)^k = I - (I - A)^{N+1} \to I$$

as $N \to \infty$, we conclude that $C = A^{-1}$. □

Exercise. Show that the preceding result is equivalent to the following statement: If $\|\cdot\|$ is a matrix norm, and if $\|A\| < 1$, then $I - A$ is invertible and

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k$$

Exercise. Let $\|\cdot\|$ be a matrix norm on M_n, and suppose a given matrix $A \in M_n$ has an "approximate inverse" $B \in M_n$ with the property that $\|BA - I\| < 1$. Show that A and B are both invertible.

Exercise. If the matrix norm $\|\cdot\|$ has the property that $\|I\| = 1$ (which would be the case if it were an induced norm), and if $A \in M_n$ is such that $\|A\| < 1$, show that

$$\frac{1}{1 + \|A\|} \le \|(I - A)^{-1}\| \le \frac{1}{1 - \|A\|}$$

Hint: Use the inequality $\|(I - A)^{-1}\| \le \sum_{k=0}^{\infty} \|A\|^k$ to get the upper bound. Use the general inequality $\|B^{-1}\| \ge 1/\|B\|$ and the triangle inequality for the lower bound.

Exercise. If $\|\cdot\|$ is a general matrix norm, all we know is that $\|I\| \ge 1$. In this case, show that

$$\frac{\|I\|}{\|I\|+\|A\|} \leq \|(I-A)^{-1}\| \leq \frac{\|I\|-(\|I\|-1)\|A\|}{1-\|A\|}$$

whenever $\|A\|<1$.

Exercise. If $A, B \in M_n$, if A is invertible, and if $A+B$ is singular, show that $\|B\| \geq 1/\|A^{-1}\|$ for any matrix norm $\|\cdot\|$. Thus, there is an intrinsic limit to how well a nonsingular matrix can be approximated by a singular one. *Hint:* $A+B=A(I+A^{-1}B)$. If $\|A^{-1}B\|<1$, then $I+A^{-1}B$ would be invertible, so it must be that $\|A^{-1}B\| \geq 1$.

One useful and easily computed criterion for invertibility follows easily from the last corollary.

5.6.17 **Corollary.** Let $A=[a_{ij}] \in M_n$, and suppose that

$$|a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \quad \text{for all} \quad i=1,2,\ldots,n$$

Then A is invertible.

Proof: The hypothesis ensures that all main diagonal entries a_{ii} are nonzero. Set $D=\operatorname{diag}(a_{11},\ldots,a_{nn})$, so that D is an invertible diagonal matrix, $D^{-1}A$ has all 1's on the main diagonal, the matrix $B=[b_{ij}]=I-D^{-1}A$ has all 0's on the main diagonal, and $b_{ij}=-a_{ij}/a_{ii}$ if $i\neq j$. Consider the maximum row sum norm $\|\cdot\|_\infty$. The hypothesis guarantees that $\|B\|_\infty<1$, so $I-B=D^{-1}A$ is invertible by (5.6.16), and hence A is invertible. □

A matrix that satisfies the hypothesis of (5.6.17) is said to be *strictly diagonally dominant*. This sufficient condition for invertibility is known as the Levy–Desplanques theorem, and it can be improved somewhat. See Sections (6.1), (6.2), and (6.4).

We now consider in more detail the induced matrix norms defined in (5.6.1). These are some of the most familiar matrix norms, and they have an important minimality property. Because one often wishes to establish that a given matrix A is convergent by using the test $\|A\|<1$, it is natural to prefer matrix norms that are uniformly as small as possible. As we shall show, the entire class of induced matrix norms has this desirable property, and this property characterizes the class of induced matrix norms.

Any two norms on a finite-dimensional space are equivalent, and so for each two matrix norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ there is a least finite positive constant $C_M(\alpha,\beta)$ such that $\|A\|_\alpha \leq C_M(\alpha,\beta)\|A\|_\beta$ for all $A\in M_n$. This constant can be computed as

$$C_M(\alpha,\beta)=\max_{A\neq 0}\frac{\|A\|_\alpha}{\|A\|_\beta}$$

If the roles of α and β are reversed, there must be a similarly defined least finite positive constant $C_M(\beta,\alpha)$ such that $\|A\|_\beta \leq C_M(\beta,\alpha)\|A\|_\alpha$ for all $A\in M_n$. In general, there is no obvious relation between the two constants $C_M(\alpha,\beta)$ and $C_M(\beta,\alpha)$, but if we examine the table in Problem 23 at the end of this section, we see that its upper left 3×3 corner is symmetric; that is, $C_M(\alpha,\beta)=C_M(\beta,\alpha)$ for any pair of the three matrix norms $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$. All three of these matrix norms are induced norms, and this symmetry is a property of all induced norms.

5.6.18 **Theorem.** Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be two given vector norms on $\mathbf{C}^n$, and let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ denote the respective induced matrix norms on M_n, that is,

$$\|A\|_\alpha \equiv \max_{x\neq 0}\frac{\|Ax\|_\alpha}{\|x\|_\alpha} \quad\text{and}\quad \|A\|_\beta \equiv \max_{x\neq 0}\frac{\|Ax\|_\beta}{\|x\|_\beta}$$

Define

$$R_{\alpha\beta} \equiv \max_{x\neq 0}\frac{\|x\|_\alpha}{\|x\|_\beta} \quad\text{and}\quad R_{\beta\alpha} \equiv \max_{x\neq 0}\frac{\|x\|_\beta}{\|x\|_\alpha} \tag{5.6.19}$$

Then

$$\max_{A\neq 0}\frac{\|A\|_\alpha}{\|A\|_\beta} = R_{\alpha\beta}R_{\beta\alpha} \tag{5.6.20}$$

In particular,

$$\max_{A\neq 0}\frac{\|A\|_\alpha}{\|A\|_\beta} = \max_{A\neq 0}\frac{\|A\|_\beta}{\|A\|_\alpha} = R_{\alpha\beta}R_{\beta\alpha} \tag{5.6.21}$$

Proof: Let $A\in M_n$ and $x\in\mathbf{C}^n$ be given, and suppose that $x\neq 0$ and $Ax\neq 0$. Then

$$\frac{\|Ax\|_\alpha}{\|x\|_\alpha} = \frac{\|Ax\|_\alpha}{\|Ax\|_\beta}\frac{\|Ax\|_\beta}{\|x\|_\beta}\frac{\|x\|_\beta}{\|x\|_\alpha} \leq R_{\alpha\beta}\frac{\|Ax\|_\beta}{\|x\|_\beta}R_{\beta\alpha}$$

an inequality that holds even if $Ax=0$. Thus,

$$\|A\|_\alpha \equiv \max_{x\neq 0}\frac{\|Ax\|_\alpha}{\|x\|_\alpha} \leq R_{\alpha\beta}\max_{x\neq 0}\frac{\|Ax\|_\beta}{\|x\|_\beta}R_{\beta\alpha} \equiv R_{\alpha\beta}R_{\beta\alpha}\|A\|_\beta$$

and hence

$$\frac{\|A\|_\alpha}{\|A\|_\beta} \le R_{\alpha\beta} R_{\beta\alpha} \tag{5.6.22}$$

for all nonzero $A \in M_n$.

Each of the two extrema in (5.6.19) is achieved for some nonzero vector, so there are vectors $y, z \in \mathbf{C}^n$ such that $\|y\|_2 = \|z\|_2 = 1$, $\|y\|_\alpha = R_{\alpha\beta}\|y\|_\beta$, and $\|z\|_\beta = R_{\beta\alpha}\|z\|_\alpha$. By Corollary (5.5.15) there exists a vector $z_0 \in \mathbf{C}^n$ such that

(a) $|z_0^* x| \le \|x\|_\beta$ for all $x \in \mathbf{C}^n$; and
(b) $z_0^* z = \|z\|_\beta$.

Consider the matrix $A_0 \equiv y z_0^*$. Using (b), we have

$$\frac{\|A_0 z\|_\alpha}{\|z\|_\alpha} = \frac{\|y z_0^* z\|_\alpha}{\|z\|_\alpha} = \frac{\|y\|_\alpha |z_0^* z|}{\|z\|_\alpha} = \frac{\|y\|_\alpha \|z\|_\beta}{\|z\|_\alpha}$$

so we have the lower bound

$$\|A_0\|_\alpha \ge \frac{\|y\|_\alpha \|z\|_\beta}{\|z\|_\alpha} = R_{\alpha\beta} R_{\beta\alpha} \|y\|_\beta$$

On the other hand, we can use (a) to obtain

$$\frac{\|A_0 x\|_\beta}{\|x\|_\beta} = \frac{\|y z_0^* x\|_\beta}{\|x\|_\beta} = \frac{\|y\|_\beta |z_0^* x|}{\|x\|_\beta} \le \frac{\|y\|_\beta \|x\|_\beta}{\|x\|_\beta} = \|y\|_\beta$$

and hence we have the upper bound

$$\|A_0\|_\beta \le \|y\|_\beta$$

Combining these two bounds, we have

$$\frac{\|A_0\|_\alpha}{\|A_0\|_\beta} \ge \frac{R_{\alpha\beta} R_{\beta\alpha} \|y\|_\beta}{\|y\|_\beta} = R_{\alpha\beta} R_{\beta\alpha}$$

which shows that equality is possible in (5.6.22) and establishes (5.6.20). The assertion (5.6.21) follows because the right-hand side of the identity (5.6.20) is symmetric in α and β. □

Is it possible that two different vector norms on $\mathbf{C}^n$ could induce the same matrix norm on M_n? According to the following consequence of (5.6.18), this can happen if and only if one of the vector norms is a constant scalar multiple of the other.

5.6.23 Corollary. Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be vector norms on $\mathbf{C}^n$, and let $|||\cdot|||_\alpha$ and $|||\cdot|||_\beta$ denote the respective induced matrix norms on M_n. Then $|||A|||_\alpha = |||A|||_\beta$ for all $A \in M_n$ if and only if there is a positive constant c such that $\|x\|_\alpha = c\|x\|_\beta$ for all $x \in \mathbf{C}^n$.

Proof: Observe that

$$R_{\beta\alpha} = \max_{x \ne 0} \frac{\|x\|_\beta}{\|x\|_\alpha} = \left[\min_{x \ne 0} \frac{\|x\|_\alpha}{\|x\|_\beta}\right]^{-1} \ge \left[\max_{x \ne 0} \frac{\|x\|_\alpha}{\|x\|_\beta}\right]^{-1} = \frac{1}{R_{\alpha\beta}}$$

Thus, we have the general inequality

$$R_{\alpha\beta} R_{\beta\alpha} \ge 1 \tag{5.6.24}$$

with equality if and only if

$$\min_{x \ne 0} \frac{\|x\|_\alpha}{\|x\|_\beta} = \max_{x \ne 0} \frac{\|x\|_\alpha}{\|x\|_\beta}$$

which can occur if and only if the function $\|x\|_\alpha / \|x\|_\beta$ is constant for all $x \ne 0$. Thus, if $\|x\|_\alpha \equiv c\|x\|_\beta$, we certainly have $R_{\alpha\beta} R_{\beta\alpha} = 1$ and hence $|||A|||_\alpha \le |||A|||_\beta$ and $|||A|||_\beta \le |||A|||_\alpha$ for all $A \in M_n$ by (5.6.21); in this event, $|||A|||_\alpha = |||A|||_\beta$ for all $A \in M_n$. Conversely, if the two induced matrix norms are identical, then $R_{\alpha\beta} R_{\beta\alpha} = 1$ by (5.6.20) and hence equality holds in (5.6.24) and the ratio $\|x\|_\alpha / \|x\|_\beta$ is constant by the preceding argument. □

5.6.25 Corollary. Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be vector norms on $\mathbf{C}^n$, and let $|||\cdot|||_\alpha$ and $|||\cdot|||_\beta$ denote the respective induced matrix norms on M_n. Then $|||A|||_\alpha \le |||A|||_\beta$ for all $A \in M_n$ if and only if $|||A|||_\alpha = |||A|||_\beta$ for all $A \in M_n$.

Proof: If $|||A|||_\alpha \le |||A|||_\beta$ for all $A \in M_n$, then $R_{\alpha\beta} R_{\beta\alpha} \le 1$, which [because of (5.6.24)] implies that $R_{\alpha\beta} R_{\beta\alpha} = 1$. Therefore, $|||A|||_\alpha \le |||A|||_\beta$ and $|||A|||_\beta \le |||A|||_\alpha$ for all $A \in M_n$ by (5.6.21). □

The last corollary says that no induced matrix norm can be uniformly dominated by another. What happens if we permit comparisons with other (not necessarily induced) matrix norms?

5.6.26 Theorem. Let $|||\cdot|||$ be a given matrix norm on M_n, and let $|||\cdot|||_\alpha$ be a given induced matrix norm on M_n. Then

(a) There is an induced matrix norm $N(\cdot)$ on M_n such that $N(A) \le |||A|||$ for every $A \in M_n$; and
(b) $|||A||| \le |||A|||_\alpha$ for every $A \in M_n$ if and only if $|||A||| = |||A|||_\alpha$ for every $A \in M_n$.

Proof: Define the vector norm $\|\cdot\|$ on $\mathbf{C}^n$ by

$$\|x\| \equiv |||X|||, \qquad X \equiv [x\ x \ldots x] \in M_n \tag{5.6.27}$$

and consider the matrix norm $N(\cdot)$ on M_n that is induced by $\|\cdot\|$. For any $A \in M_n$, we have

$$\begin{aligned} N(A) &\equiv \max_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \max_{x \neq 0} \frac{\|[Ax\ Ax \dots Ax]\|}{\|[x\ x \dots x]\|} \\ &= \max_{x\neq 0} \frac{\|AX\|}{\|X\|} \qquad (5.6.28) \\ &\leq \max_{x \neq 0} \frac{\|A\|\,\|X\|}{\|X\|} \qquad \text{(because } \|\cdot\| \text{ is a matrix norm)} \\ &= \|A\| \end{aligned}$$

which establishes (a). To prove (b), suppose that $\|A\| \leq \|A\|_\alpha$ for all $A \in M_n$. Then by (a) we have

$$N(A) \leq \|A\| \leq \|A\|_\alpha$$

for all $A \in M_n$. But $N(\cdot)$ and $\|\cdot\|_\alpha$ are both induced norms, so $N(A) \equiv \|A\|_\alpha$ by (5.6.25), and hence $\|A\| \equiv \|A\|_\alpha$ for all $A \in M_n$. □

The preceding result is the motivation for the following definition.

5.6.29 **Definition.** A matrix norm $\|\cdot\|$ on M_n is a *minimal matrix norm* if the only matrix norm $N(\cdot)$ on M_n such that $N(A) \leq \|A\|$ for all $A \in M_n$ is $N(\cdot) = \|\cdot\|$.

Assertion (b) of Theorem (5.6.26) says that every induced norm on M_n is minimal. Assertion (a) implies immediately that every minimal norm is induced. Thus, if one wants to use a matrix norm that cannot be uniformly improved upon (in terms of small values on all matrices), one should use an induced norm, and any norm with this optimality property must be an induced norm.

The vector norm (5.6.27) is a special case of a whole family of vector norms that can be constructed from a given matrix norm. Let $\|\cdot\|$ be a given matrix norm on M_n, let $y \in \mathbf{C}^n$ be a given nonzero vector, and define the function $\|\cdot\|_y : \mathbf{C}^n \to \mathbf{R}$ by

$$\|x\|_y \equiv \|xy^*\| \qquad y \in \mathbf{C}^n, \quad y \neq 0 \qquad (5.6.30)$$

Then $\|\cdot\|_y$ is a vector norm on $\mathbf{C}^n$ with the property that

$$\|Ax\|_y = \|A(xy^*)\| \leq \|A\|\,\|xy^*\| = \|A\|\,\|x\|_y$$

for all $A \in M_n$. If $y = [1 1 \dots 1]^T$, then (5.6.30) reduces to (5.6.27). If we denote by $N_y(\cdot)$ the matrix norm on M_n that is induced by $\|\cdot\|_y$, this inequality says that

$$N_y(A) \equiv \max_{x\neq 0} \frac{\|Ax\|_y}{\|x\|_y} \leq \max_{x \neq 0} \frac{\|A\|\,\|x\|_y}{\|x\|_y} = \|A\| \quad \text{for all} \quad A \in M_n \qquad (5.6.31)$$

This is evidently a generalization of (5.6.26a).

If the given matrix norm $\|\cdot\|$ is a minimal norm, then (5.6.31) implies that $\|A\| = N_y(A)$ for all $A \in M_n$. Since the vector y used in this argument can be *any* nonzero vector, we would then have $N_y(\cdot) = \|\cdot\| = N_z(\cdot)$ for all nonzero $y, z \in M_n$.

5.6.32 **Theorem.** Let $\|\cdot\|$ be a matrix norm on M_n, and let $N_y(\cdot)$ be the induced norm defined by (5.6.31) and (5.6.30). The following are equivalent:

(a) $\|\cdot\|$ is an induced matrix norm.
(b) $\|\cdot\|$ is a minimal matrix norm.
(c) $\|\cdot\| = N_y(\cdot)$ for all nonzero $y \in \mathbf{C}^n$.

Proof: The assertion that (a) implies (b) is just (5.6.26b). We have just observed that if $\|\cdot\|$ is minimal, then $\|\cdot\| = N_y(\cdot)$, so (b) implies (c). If (c), then $\|\cdot\|$ is induced because $N_y(\cdot)$ is induced by definition. □

There is somewhat more to be gleaned from these observations. If $N_y(\cdot) = \|\cdot\|$ for all nonzero $y \in \mathbf{C}^n$, then $N_y(\cdot) = N_z(\cdot)$ for all nonzero $y, z \in \mathbf{C}^n$. But Corollary (5.6.23) says that the vector norm that induces a given matrix norm is unique up to a scale factor, so $\|\cdot\|_y = c_{yz}\|\cdot\|_z$ for some positive constant c_{yz}.

Exercise. If the matrix norm $\|\cdot\|$ on M_n is induced by the vector norm $\|\cdot\|$ on $\mathbf{C}^n$, show that $\|yz^*\| = \|y\|\,\|z\|^D$, $\|\cdot\|_z = \|\cdot\|\,\|z\|^D$, and $c_{yz} = \|y\|^D/\|z\|^D$ for all $y, z \in \mathbf{C}^n$. The vector norm $\|\cdot\|^D$ is the dual of the vector norm $\|\cdot\|$, as defined in (5.4.12).

5.6.33 **Theorem.** Let $\|\cdot\|$ be a given matrix norm on M_n and let $\|\cdot\|_y$ be the vector norm on $\mathbf{C}^n$ defined by (5.6.30). The following two assertions are equivalent:

(a) For each pair of nonzero vectors $y, z \in \mathbf{C}^n$ there is a positive constant c_{yz} such that

$$\|x\|_y = c_{yz}\|x\|_z \quad \text{for all} \quad x \in \mathbf{C}^n$$

(b) $$\|xy^*\| = \frac{\|xz^*\|\,\|zy^*\|}{\|zz^*\|} \quad \text{for all} \quad x, y, z \in \mathbf{C}^n \quad \text{with} \quad z \neq 0$$

If $\|\cdot\|$ is an induced matrix norm, then it satisfies the identity (b), and the vector norms constructed from it by (5.6.30) satisfy (a).

Proof: If (a), then

$$\|xz^*\|\,\|zy^*\| = \|x\|_z\|z\|_y = (1/c_{yz})\|x\|_y c_{yz}\|z\|_z = \|x\|_y\|z\|_z = \|xy^*\|\,\|zz^*\|$$

Conversely, if (b), then (a) follows with $c_{yz} = \|zy^*\|/\|zz^*\|$. We have already argued that if $N_y(\cdot) = \|\cdot\|$, then (a) [and hence (b) also] must follow, and this will be the case if $\|\cdot\|$ is an induced norm by (5.6.32). □

Exercise. Any positive scalar multiple of an induced norm satisfies the identity (5.6.33b). Show that the matrix norms $\|\cdot\|_1$ and $\|\cdot\|_2$ both satisfy this identity, but that neither norm is a scalar multiple of an induced norm.

We saw in (5.6.2) that if $\|\cdot\|$ is an induced matrix norm, then $\|I\| = 1$. This property is unfortunately not sufficient for a matrix norm to be an induced norm. It is easy to show that the function

$$\|A\| \equiv \max\{\|A\|_1, \|A\|_\infty\} \tag{5.6.34}$$

defines a matrix norm on M_n, and that $\|I\| = 1$. But since $\|A\|_1 \le \|A\|$ for all $A \in M_n$ and $\|A\|_1 < \|A\|$ for $A = \left[\begin{smallmatrix}1 & 0\\ 1 & 3\end{smallmatrix}\right]$, $\|\cdot\|$ is not a minimal norm and hence cannot be an induced norm.

Exercise. Verify that (5.6.34) defines a matrix norm. More generally, show that if $\|\cdot\|_{(1)}, \ldots, \|\cdot\|_{(k)}$ are given matrix norms on M_n, then

$$\|A\| \equiv \max\{\|A\|_{(1)}, \ldots, \|A\|_{(k)}\}$$

defines a matrix norm on M_n.

The induced norms are minimal among all matrix norms, but suppose one considers only the important class of *unitarily invariant matrix norms*. These are the matrix norms $\|\cdot\|$ such that $\|A\| = \|UAV\|$ for all $A \in M_n$ and all unitary matrices $U, V \in M_n$. It turns out that in this class there is only one minimal matrix norm, and that is the spectral norm.

5.6.35 **Corollary.** If $\|\cdot\|$ is a unitarily invariant matrix norm, then $\|A\|_2 \le \|A\|$ for all $A \in M_n$. The spectral norm $\|\cdot\|_2$ is the only matrix norm on M_n that is both induced and unitarily invariant.

Proof: Suppose that $\|\cdot\|$ is a given unitarily invariant matrix norm. By part (a) of Theorem (5.6.26), we know that $N(A) \le \|A\|$ for all $A \in M_n$,

where $N(A)$ is induced by the vector norm $\|\cdot\|$ defined by (5.6.27). If $U \in M_n$ is unitary, we have $\|Ux\| = \|UX\| = \|X\| = \|x\|$, and hence the vector norm $\|\cdot\|$ is unitarily invariant. If $x \in \mathbf{C}^n$ is a given nonzero vector, there exists a unitary matrix U such that $Ux = \|x\|_2 e_1$. Thus, $\|x\| = \|\|x\|_2 U^* e_1\| = \|x\|_2 \|U^* e_1\| = \|x\|_2\|e_1\|$ for all $x \in \mathbf{C}^n$. The vector norm $\|\cdot\|$ is therefore a scalar multiple of the Euclidean norm and Corollary (5.6.23) says that $N(\cdot)$ (the matrix norm induced by $\|\cdot\|$) equals $\|\cdot\|_2$ (the matrix norm induced by $\|\cdot\|_2$). Therefore, $\|\cdot\|_2 = N(A) \le \|A\|$ for all $A \in M_n$. If $\|\cdot\|$ is assumed to be induced, then it is minimal and hence $\|A\|_2 = \|A\|$ for all $A \in M_n$. □

If $\|\cdot\|$ is a matrix norm on M_n, then the function $\|\cdot\|^*$ defined by

$$\|A\|^* \equiv \|A^*\|$$

is also a matrix norm on M_n. A direct calculation shows that $\|A\|_2^* = \|A^*\|_2 = \|A\|_2$ and $\|A\|_1^* = \|A^*\|_1 = \|A\|_1$ for all $A \in M_n$, but not every matrix norm has this property since $\|A\|_1^* = \|A\|_\infty \ne \|A\|_1$. A matrix norm such that $\|\cdot\|^* \equiv \|\cdot\|$ is said to be *self-adjoint*. The Frobenius and l_1 matrix norms are self-adjoint, and since

$$\|A^*\|_2^2 = \rho(AA^*) = \rho(A^*A) = \|A\|_2^2$$

the spectral norm is self-adjoint, too. In fact, all unitarily invariant norms on M_n are self-adjoint [see (7.4), Problem 2]. The spectral norm is distinguished as the only induced matrix norm that is self-adjoint.

5.6.36 **Theorem.** Let $\|\cdot\|$ be a given matrix norm on M_n. Then

(a) $\|\cdot\|^*$ is an induced norm if and only if $\|\cdot\|$ is an induced norm.
(b) If the matrix norm $\|\cdot\|$ is induced by the vector norm $\|\cdot\|$, then $\|\cdot\|^*$ is induced by the dual norm $\|\cdot\|^D$.
(c) The spectral norm $\|\cdot\|_2$ is the only matrix norm on M_n that is both induced and self-adjoint.

Proof: If $N(\cdot)$ is a matrix norm, and if $N(A) \le \|A\|^* = \|A^*\|$ for all $A \in M_n$, then $N(A)^* = N(A^*) \le \|A\|$ for all $A \in M_n$. If $\|\cdot\|$ is a minimal matrix norm, then $N(\cdot)^* = \|\cdot\|$ and hence $N(\cdot) = \|\cdot\|^*$, so $\|\cdot\|^*$ is a minimal matrix norm. The assertion (a) follows from (5.6.32). Now suppose that $\|\cdot\|$ is induced by the vector norm $\|\cdot\|$. Using the duality theorem (5.5.14), we have

$$\|A\|^* = \|A^*\| = \max_{\|x\|=1} \|A^*x\| = \max_{\|x\|=1} (\|A^*x\|^D)^D$$

$$= \max_{\|x\|=1}\ \max_{\|z\|^D=1} |(A^*x)^*z| = \max_{\|z\|^D=1}\ \max_{\|x\|=1} |x^*Az|$$

$$= \max_{\|z\|^D=1} \|Az\|^D$$

and hence $\|\cdot\|^*$ is induced by $\|\cdot\|^D$. For the last assertion, we observe that if the matrix norm $\|\cdot\|$ is induced by the vector norm $\|\cdot\|$, and if $\|\cdot\| = \|\cdot\|^*$, then (b) says that $\|\cdot\|$ is also induced by $\|\cdot\|^D$. But Corollary (5.6.23) says that the vector norm that induces a given matrix norm is uniquely determined up to a positive scalar factor, and hence there exists some $c>0$ such that $\|\cdot\|^D = c\|\cdot\|$. By (5.4.16) we must then have $\|\cdot\| = \|\cdot\|_2/\sqrt{c}$. Since the given vector norm is a multiple of the Euclidean vector norm, they both induce the same matrix norm and we conclude that $\|\cdot\| = \|\cdot\|_2$. □

Exercise. Show that $\|\cdot\|^*$ is a matrix norm whenever $\|\cdot\|$ is a matrix norm.

Exercise. Give an example to show that a self-adjoint matrix norm need not be unitarily invariant.

Absolute and monotone vector norms were introduced in (5.5), and are the most commonly used vector norms. There is a simple and useful characterization of the matrix norms that are induced by monotone vector norms.

5.6.37 **Theorem.** Let $\|\cdot\|$ be a vector norm on $\mathbf{C}^n$ and let $\|\cdot\|$ be the matrix norm on M_n that it induces. The following are equivalent:

(a) $\|\cdot\|$ is an absolute norm; that is, $\||x|\| = \|x\|$ for all $x \in \mathbf{C}^n$.
(b) $\|\cdot\|$ is a monotone norm; that is, $\|x\| \le \|y\|$ whenever $|x| \le |y|$.
(c) Whenever $D = \text{diag}(d_1, d_2, \ldots, d_n) \in M_n$, then

$$\|D\| = \max_{1\le i\le n} |d_i|$$

Proof: The equivalence of (a) and (b) is the content of (5.5.10). If $\|\cdot\|$ is monotone, and if we set

$$d \equiv \max_{1\le i\le n} |d_i|, \qquad d = |d_k|$$

then $|Dx| \le |dx|$ and hence $\|Dx\| \le d\|x\|$ with equality for $x = e_k$. Thus,

$$\|D\| = \max_{x\ne 0} \frac{\|Dx\|}{\|x\|} = d$$

and hence (b) implies (c). If we assume (c), let $x, y \in \mathbf{C}^n$ be given with $|x| \le |y|$ and note that there are complex numbers d_k such that $|x_k| = d_k y_k$ and $|d_k| \le 1$, $k = 1, \ldots, n$. Thus, if $D \equiv \text{diag}(d_1, \ldots, d_n)$, we have $Dy = |x|$ and $\|D\| \le 1$. Since

$$\||x|\| = \|Dy\| \le \|D\|\,\|y\| \le \|y\|$$

the norm $\|\cdot\|$ must be monotone. □

Problems

1. Give an example of a vector norm for matrices for which $\|I\| < 1$.

2. A matrix A such that $A^2 = A$ is said to be *idempotent*. Give an example of a 2-by-2 idempotent matrix other than I and 0. Show that 0 and 1 are the only possible eigenvalues of an idempotent matrix. Show that an idempotent matrix A must always be diagonalizable and that $\|A\| \ge 1$ for any matrix norm $\|\cdot\|$ if $A \ne 0$.

3. If $\|\cdot\|$ is a matrix norm on M_n, show that $c\|\cdot\|$ is a matrix norm for all $c \ge 1$. Show, however, that neither $c\|\cdot\|_1$ nor $c\|\cdot\|_1$ is a matrix norm for any $c < 1$.

4. In Definition (5.6.1) the same vector norm is involved in two different ways. More generally, we might define $\|\cdot\|_{\alpha,\beta}$ by

$$\|A\|_{\alpha,\beta} \equiv \max_{\|x\|_\alpha=1} \|Ax\|_\beta$$

where $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are two (possibly different) vector norms. Is such a function $\|\cdot\|_{\alpha,\beta}$ a matrix norm? Study $\|\cdot\|_{\alpha,\beta}$ to determine what interesting properties it might have; note that this notion might be used to define a norm on m-by-n matrices, since $\|\cdot\|_\alpha$ may be taken to be a vector norm on $\mathbf{C}^m$ and $\|\cdot\|_\beta$ may be taken to be a vector norm on $\mathbf{C}^n$. What properties like those of an induced matrix norm does $\|\cdot\|_{\alpha,\beta}$ have in this regard?

5. Show that both the Euclidean norm $\|\cdot\|_2$ and the spectral norm $\|\cdot\|_2$ are unitarily invariant norms on M_n; that is, A and UAV have the same norm whenever U and V are unitary. Compare the matrix norms $\|\cdot\|_2$ and $\|\cdot\|_2$ in as many respects as you can. Note that $\|A\|_2 = [\text{tr}\, A^*A]^{1/2}$.

6. Verify that axioms (1)–(3) for $\|\cdot\|$ imply that the same axioms hold for $\|\cdot\|_S$ in (5.6.7). This verifies that (5.6.7) remains valid if "matrix norm" in the hypothesis and conclusion is replaced by "vector norm on matrices."

7. If $\|\cdot\|$ is an induced matrix norm on M_n and if $S \in M_n$ is nonsingular, show that $\|\cdot\|_S$ [as defined in (5.6.7)] is also an induced matrix norm. If

$\|\cdot\|$ is induced by the vector norm $\|\cdot\|$, show that the matrix norm $\|\cdot\|_S$ is induced by the vector norm $\|\cdot\|_{S^{-1}}$ [as defined in (5.3.2)].

8. Show that the nonsingular matrices of M_n are *dense* in M_n; that is, show that every matrix in M_n is the limit of nonsingular matrices. Are the singular matrices dense in M_n, too?

9. Show that the set of vector norms on $\mathbf{C}^m$ is convex for all $m \geq 1$, but the set of matrix norms on M_n is not convex for any $n \geq 2$. If $N_1(\cdot)$ and $N_2(\cdot)$ are matrix norms on M_n, show that $N(\cdot) \equiv \frac{1}{2}[N_1(\cdot)+N_2(\cdot)]$ is a matrix norm if and only if

$$[N_1(A)-N_2(A)][N_1(B)-N_2(B)] \leq 2[N_1(A)N_1(B)-N_1(AB)] + 2[N_2(A)N_2(B)-N_2(AB)]$$

for all $A, B \in M_n$. *Hint:* Consider $N_1(\cdot)=\|\cdot\|_1$, $N_2(\cdot)=\|\cdot\|_2$, $A=\begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$, and $B=A^T$. See Example (7.4.54) for an important subset of the matrix norms that *is* convex.

10. Show that the l_1 vector norm on M_n, $\|A\|_1 \equiv \sum_{i,j=1}^n |a_{ij}|$, is a matrix norm that is not an induced norm.

11. Show that all of the following are equivalent ways to compute the spectral norm (5.6.6):

$$\begin{aligned} \|A\|_2 &= \max_{\|x\|_2=1} \|Ax\|_2 = \max_{\|x\|_2 \leq 1} \|Ax\|_2 \\ &= \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{\|x\|_2=\|y\|_2=1} |y^*Ax| \\ &= \max_{\substack{\|x\|_2 \leq 1 \\ \|y\|_2 \leq 1}} |y^*Ax| \end{aligned}$$

Use these identities to show that $\|A\|_2 = \|A^*\|_2$ for all $A \in M_n$. Now prove that $\|AA^*\|_2 = \|A^*A\|_2 = \|A\|_2^2$ by using the facts that $\|\cdot\|_2$ is a matrix norm and A^*A is Hermitian.

12. If $\rho(A)<1$, $A \in M_n$, show that the series $I+A+A^2+\cdots$ converges to the sum $(I-A)^{-1}$.

13. If $A \in M_n$ is not invertible, show that $\|I-A\| \geq 1$ for every matrix norm $\|\cdot\|$.

14. Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be given matrix norms on M_n. Show that $\|A\| \equiv \max\{\|A\|_\alpha, \|A\|_\beta\}$ is a matrix norm on M_n. When is it an induced norm?

15. Give an example of a matrix A such that $\rho(A) < \|A\|$ for every matrix norm $\|\cdot\|$.

16. Let $A=[a_{ij}] \in M_n$. Show that the function $\|\cdot\|$ defined on M_n by $\|A\| \equiv n \max_{1 \leq i,j \leq n} |a_{ij}|$ is a matrix norm that is not induced when $n \geq 2$.

17. Use the idea in Problem 12 to compute the inverse of the matrix

$$\begin{bmatrix} 1 & -2 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{bmatrix}$$

Hint: Only three terms in the series are nonzero.

18. Explain how to generalize the method in Problem 17 to invert a general nonsingular upper triangular matrix $A \in M_n$. *Hint:* Choose a diagonal matrix D such that DA has all 1's on the main diagonal.

19. Show that the spectral radius $\rho(\cdot)$ is a continuous and homogeneous function on M_n, but that it is neither a matrix norm nor a vector norm on M_n because

(a) $\rho(A)=0$ is possible for some $A \neq 0$;
(b) $\rho(A+B) > \rho(A)+\rho(B)$ is possible; and
(c) $\rho(AB) > \rho(A)\rho(B)$ is possible even if $\rho(A)$ and $\rho(B)$ are both nonzero.

Hint: Consider $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, $\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$, $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, and $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$.

20. Show that $\|AB\|_2 \leq \|A\|_2 \|B\|_2$ and $\|AB\|_2 \leq \|A\|_2 \|B\|_2$ for all $A, B \in M_n$.

21. Show that $\|A\|_2^2 \leq \|A\|_1 \|A\|_\infty$ for all $A \in M_n$. How does this compare with the bound you can get directly from the table in Problem 23? Why the difference? *Hint:* $\rho(A^*A) \leq \|A^*A\|_1$ and $\|A^*\|_1 = \|A\|_\infty$.

22. Let $\|\cdot\|_\alpha$ be a given vector norm on $\mathbf{C}^n$ and define $\|\cdot\|_\beta \equiv (\|\cdot\|_\alpha)^D$ to be its dual norm. Let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ denote the matrix norms on M_n that are induced by $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$, respectively. Use (5.6.36) to show that $\|A^*\|_\beta = \|A\|_\alpha$ for all $A \in M_n$. Deduce that $\|A\|_2^2 \leq \|A\|_\alpha \|A\|_\beta$ for all $A \in M_n$ and explain how this generalizes the result in Problem 21. How is this inequality related to (5.4.13) for $x=y$?

23. Verify that the entries in the following table give the best constants C_M such that $\|A\|_\alpha \leq C_M \|A\|_\beta$ for all $A \in M_n$. All the norms in the table are matrix norms. The hints (i,j) following the table pertain to establishing the (i,j) entries in the table. The matrix given in each case is one for which the inequality $\|A\|_\alpha \leq C_M \|A\|_\beta$ is an equality for the given value of the constant C_M.

$\lVert\cdot\rVert_\alpha$ \ $\lVert\cdot\rVert_\beta$	$\lVert\cdot\rVert_1$	$\lVert\cdot\rVert_2$	$\lVert\cdot\rVert_\infty$	$\lvert\cdot\rvert_1$	$\lvert\cdot\rvert_2$	$n\lvert\cdot\rvert_\infty$
$\lVert\cdot\rVert_1$	1	$\sqrt{n}$	n	1	$\sqrt{n}$	1
$\lVert\cdot\rVert_2$	$\sqrt{n}$	1	$\sqrt{n}$	1	1	1
$\lVert\cdot\rVert_\infty$	n	$\sqrt{n}$	1	1	$\sqrt{n}$	1
$\lvert\cdot\rvert_1$	n	$n^{3/2}$	n	1	n	n
$\lvert\cdot\rvert_2$	$\sqrt{n}$	$\sqrt{n}$	$\sqrt{n}$	1	1	1
$n\lvert\cdot\rvert_\infty$	n	n	n	n	n	1

The following matrices are all in M_n:

I is the identity matrix

J has all entries 1

A_1 has all 1's in its first column and all other entries are 0

A_2 has only the (1, 1) entry 1 and all other entries are 0

(1, 2) follows from (2, 1) by (5.6.21)

(1, 3) $\lVert A\rVert_1 \le \lvert A\rvert_1 \le n\lVert A\rVert_\infty$; A_1

(1, 4) A_1

(1, 5) $\max_{1\le j\le n}\left[\sum_{i=1}^n |a_{ij}|\right]^2 \le \sum_{j=1}^n\left[\sum_{i=1}^n |a_{ij}|\right]^2 \le \left[\sum_{j=1}^n 1\right]\left[\sum_{i=1}^n |a_{ij}|^2\right]$

(Cauchy–Schwarz inequality); A_1

(1, 6) $\max_{1\le j\le n}\sum_{i=1}^n |a_{ij}| \le n \max_{1\le i,j\le n}|a_{ij}|$; J

(2, 1) follows from (2, 5) and (5, 1); A_1^*

(2, 3) follows from (2, 5) and (5, 3); A_1

(2, 4) follows from (2, 5) and (5, 4); A_2

(2, 5) $\lVert A\rVert_2^2 = \rho(A^*A) \le \sum_{i=1}^n \lambda_i(A^*A) = \operatorname{tr} A^*A = \lvert A\rvert_2^2$; A_1

(2, 6) follows from (2, 5) and (5, 6); J

(3, 1) follows from (1, 3) by (5.6.21); A_1^*

(3, 2) follows from (2, 3) by (5.6.21); A_1^*

(3, 4) A_1^*

(3, 5) similar to (1, 5); A_1^*

(3, 6) similar to (1, 6); J

(4, 1) $\sum_{j=1}^n\sum_{i=1}^n |a_{ij}| \le n \max_{1\le j\le n}\sum_{i=1}^n |a_{ij}|$; I

(4, 2) follows from (4, 5) and (5, 2); the following matrix gives equality in both: take $a \equiv e^{2\pi i/n}$, notice that $(\bar{a})^k = a^{-k}$ and $\sum_{k=0}^{n-1} a^{kj} = 0$ if $j\ne 0$, n if $j=0$; let the (k, j) entry of A be a^{kj} and check that $A^*A = nI$, $\lVert A\rVert_2 = \sqrt{n}$, $\lvert A\rvert_1 = n^2$, and $\lvert A\rvert_2 = n$

(4, 3) similar to (4, 1); I

(4, 5) $\left[\sum_{i,j=1}^n |a_{ij}|\right]^2 = \sum_{i,j,p,q=1}^n |a_{ij}|\,|a_{pq}| \le \frac{1}{2}\sum_{i,j,p,q=1}^n [|a_{ij}|^2 + |a_{pq}|^2]$

(arithmetic–geometric mean inequality); J

(4, 6) $\sum_{i,j=1}^n |a_{ij}| \le n^2 \max_{1\le i,j\le n}|a_{ij}|$; J

(5, 1) $\sum_{j=1}^n\sum_{i=1}^n |a_{ij}|^2 \le \sum_{j=1}^n\left[\sum_{i=1}^n |a_{ij}|\right]^2 \le n\left[\max_{1\le j\le n}\sum_{i=1}^n |a_{ij}|\right]^2$; I

(5, 2) $\sum_{i,j=1}^n |a_{ij}|^2 = \operatorname{tr} A^*A = \sum_{i=1}^n \lambda_i(A^*A) \le n\,\lambda_{\max}(A^*A)$; I

(5, 3) similar to (5, 1); I

(5, 4) $\sum_{i,j=1}^n |a_{ij}|^2 \le \left[\sum_{i,j=1}^n |a_{ij}|\right]^2$; A_2

(5, 6) $\sum_{i,j=1}^n |a_{ij}|^2 \le n^2 \max_{1\le i,j\le n}|a_{ij}|^2$; J

(6, 1) $\max_{1\le i,j\le n}|a_{ij}| \le \max_{1\le j\le n}\sum_{i=1}^n |a_{ij}|$; I

(6, 2) $\max_{1\le i,j\le n}|a_{ij}|^2 \le \max_{1\le i\le n}\sum_{j=1}^n |a_{ij}|^2 = \max_{1\le i\le n}(A^*A)_{ii} \le \rho(A^*A)$; I

(6, 3) similar to (6, 1); I

(6, 4) $\max_{1\le i,j\le n}|a_{ij}| \le \sum_{i,j=1}^n |a_{ij}|$; A_2

(6, 5) $\max_{1\le i,j\le n}|a_{ij}|^2 \le \sum_{i,j=1}^n |a_{ij}|^2$; A_2

24. Show that the bound (5, 2) in Problem 23 can be improved to $\lvert A\rvert_2 \le [\operatorname{rank} A]^{1/2}\lVert A\rVert_2$. *Hint:* rank A = number of nonzero eigenvalues of A^*A.

25. Let $A \in M_n$ be given. If $\epsilon > 0$, we know by Lemma (5.6.10) that there is some matrix norm $\|\cdot\|$ such that $\rho(A) < \|A\| < \rho(A) + \epsilon$. Show that there is a nonsingular matrix $C = C(\epsilon) \in M_n$ such that $\rho(A) < \|CAC^{-1}\|_2 < \rho(A) + \epsilon$. *Hint:* Use the same construction as in Lemma (5.6.10) and show that $\|CAC^{-1}\|_2^2 = \rho(A^*A) + O(\epsilon)$ as $\epsilon \to 0$.

26. Show that $\|A\|_2^2 \geq \sum_{i=1}^n |\lambda_i|^2$ for all $A \in M_n$ with equality if and only if A is normal. The quantity

$$\left[\|A\|_2^2 - \sum_{i=1}^n |\lambda_i|^2\right]^{1/2}$$

is sometimes called the *defect from normality* for this reason. *Hint:* Use the Schur triangularization theorem and the fact that the Frobenius norm is unitarily invariant.

27. Theorem (5.6.9) and the companion matrix can be used to give bounds for the zeroes of a polynomial with real or complex coefficients. Any polynomial $f(z)$ of degree at least 1 can be written in the form $f(z) = Cz^k p(z)$, where C is a nonzero constant,

$$p(z) = z^n + a_{n-1}z^{n-1} + a_{n-2}z^{n-2} + \cdots + a_1 z + a_0 \tag{5.6.38}$$

and $a_0 \neq 0$. The roots of $p(z) = 0$ are the nonzero roots of $f(z) = 0$, and it is these roots for which we can give various bounds. (a) Show that the characteristic polynomial of the *companion matrix*

$$C(p) \equiv \begin{bmatrix} -a_{n-1} & -a_{n-2} & \cdots & -a_1 & -a_0 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & & 1 & 0 \end{bmatrix} \tag{5.6.39}$$

is exactly $p(z)$, and hence the eigenvalues of $C(p)$ are the same as the roots of $p(z) = 0$. *Hint:* Compute $\det[zI - C(p)]$ using cofactors of the first column and use induction. (b) Use Theorem (5.6.9) to show that if $\tilde{z}$ is a root of $p(z) = 0$ and if $\|\cdot\|$ is any matrix norm on M_n, then $|\tilde{z}| \leq \|C(p)\|$. In the following, $\tilde{z}$ represents any root of $p(z) = 0$. (c) Use $\|\cdot\|_1$ to show that

$$\begin{aligned} |\tilde{z}| &\leq \max\{|a_0|, 1 + |a_1|, \ldots, 1 + |a_{n-1}|\} \\ &\leq 1 + \max\{|a_0|, |a_1|, \ldots, |a_{n-1}|\} \end{aligned} \tag{5.6.40}$$

This bound on the roots is known as *Cauchy's bound.* (d) Use $\|\cdot\|_\infty$ to show that

$$|\tilde{z}| \leq \max\{1, |a_0| + |a_1| + \cdots + |a_{n-1}|\} \leq 1 + |a_0| + |a_1| + \cdots + |a_{n-1}| \tag{5.6.41}$$

This is known as *Montel's bound.* Show that it is poorer than Cauchy's bound. (e) Use $\|\cdot\|_1$ to show that

$$|\tilde{z}| \leq (n-1) + |a_0| + |a_1| + \cdots + |a_{n-1}|$$

which is a poorer bound than (d) for all $n > 2$. (f) Use $\|\cdot\|_2$ to show that

$$|\tilde{z}| \leq [n + |a_0|^2 + |a_1|^2 + \cdots + |a_{n-1}|^2]^{1/2}$$

which is a poorer bound than Carmichael and Mason's bound (5.6.42). (g) Use $n\|\cdot\|_\infty$ to show that

$$|\tilde{z}| \leq n \max\{1, |a_0|, |a_1|, \ldots, |a_{n-1}|\}$$

which is a poorer bound than (5.6.41).

28. Using the same notation as in Problem 27, we can improve the bound in part (f). Write the companion matrix as $C(p) = S + R$, where

$$S = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \ddots & 0 & 0 \\ \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

and

$$R = \begin{bmatrix} -a_{n-1} & -a_{n-2} & \cdots & -a_1 & -a_0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

and show that $S^*R = R^*S = 0$. Show that $\|S^*S\|_2 = 1$ and $\|R^*R\|_2 = |a_0|^2 + |a_1|^2 + \cdots + |a_{n-1}|^2$. Show that

$$\begin{aligned} \|C(p)\|_2^2 &= \|C(p)^*C(p)\|_2 = \|(S+R)^*(S+R)\|_2 \\ &= \|S^*S + R^*R\|_2 \leq \|S^*S\|_2 + \|R^*R\|_2 \end{aligned}$$

and deduce *Carmichael and Mason's bound*

$$|\tilde{z}| \leq [1 + |a_0|^2 + |a_1|^2 + \cdots + |a_{n-1}|^2]^{1/2} \tag{5.6.42}$$

29. Apply the bound (5.6.41) to the polynomial

$$q(z) \equiv (z-1)\,p(z) = z^{n+1} + (a_{n-1}-1)z^n + (a_{n-2}-a_{n-1})z^{n-1} + \cdots + (a_0-a_1)z + a_0$$

and show that

$$|\tilde z| \le \max\{1,\ |a_0| + |a_0-a_1| + \cdots + |a_{n-2}-a_{n-1}| + |a_{n-1}-1|\}$$

Show that the second term in this expression is not less than 1 and deduce another bound of Montel

$$|\tilde z| \le |a_0| + |a_0-a_1| + \cdots + |a_{n-2}-a_{n-1}| + |a_{n-1}-1| \qquad (5.6.43)$$

30. Use Montel's bound (5.6.43) to prove *Kakeya's theorem:* If $f(z) = a_n z^n + a_{n-1}z^{n-1} + \cdots + a_1 z + a_0$ is a given polynomial with real nonnegative coefficients a_i that are monotone in the sense that $a_n \ge a_{n-1} \ge \cdots \ge a_1 \ge a_0$, then all the roots of $f(z)=0$ lie in the unit disc; that is, all $|\tilde z| \le 1$.

31. The preceding four problems have all concerned upper bounds on the absolute values of the roots of $p(z)=0$, but they can be used to obtain lower bounds as well. Show that if $p(z)$ is given by (5.6.38) with $a_0 \ne 0$, then the function

$$q(z) \equiv \frac{1}{a_0} z^n p\left(\frac{1}{z}\right) = z^n + \frac{a_1}{a_0} z^{n-1} + \frac{a_2}{a_0} z^{n-2} + \cdots + \frac{a_{n-1}}{a_0} z + \frac{1}{a_0}$$

is a polynomial of degree n whose zeroes are exactly the reciprocals of the roots of $p(z)=0$. Use the respective upper bounds on the roots of $q(z)=0$ to obtain the following lower bounds on the roots $\tilde z$ of $p(z)=0$.

Cauchy:

$$|\tilde z| \ge \frac{|a_0|}{\max\{1,\ |a_0|+|a_{n-1}|,\ |a_0|+|a_{n-2}|, \ldots, |a_0|+|a_1|\}} \ge \frac{|a_0|}{|a_0| + \max\{1, |a_{n-1}|, |a_{n-2}|, \ldots, |a_1|\}}$$

Montel:

$$|\tilde z| \ge \frac{|a_0|}{\max\{|a_0|,\ 1+|a_1|+|a_2|+\cdots+|a_{n-1}|\}} \ge \frac{|a_0|}{1+|a_0|+|a_1|+\cdots+|a_{n-1}|}$$

Carmichael and Mason:

$$|\tilde z| \ge \frac{|a_0|}{[1+|a_0|^2+|a_1|^2+\cdots+|a_{n-1}|^2]^{1/2}}$$

32. When the lower bounds in Problem 31 are combined with the upper bounds in Problems 27–30, it is possible to locate the zeroes of $p(z)$ in an annulus $\{z : r_1 \le |z| \le r_2\}$. As an example, consider

$$f(z) = \frac{1}{n!} z^n + \frac{1}{(n-1)!} z^{n-1} + \cdots + \frac{1}{2} z^2 + z + 1$$

which is the nth partial sum of the power series for the exponential function e^z. Show that all roots $\tilde z$ of $f(z)=0$ satisfy the inequality

$$\tfrac{1}{2} \le |\tilde z| \le 1 + n!$$

Apply Kakeya's theorem to $z^n f(1/z)$ to show that all the roots actually satisfy $|\tilde z| \ge 1$.

33. Since $\rho(A) = \rho(D^{-1}AD)$ for any nonsingular matrix D, the methods used in Problem 27 can be applied to $D^{-1}C(p)D$ to obtain other bounds on the zeroes of the polynomial $p(z)$ in (5.6.38). Make the computationally convenient choice $D = \operatorname{diag}(p_1, p_2, \ldots, p_n)$ with all $p_i > 0$ and generalize Cauchy's bound (5.6.40) to

$$|\tilde z| \le \max\left\{|a_0|\frac{p_n}{p_1},\ |a_1|\frac{p_{n-1}}{p_1} + \frac{p_{n-1}}{p_n},\ |a_2|\frac{p_{n-2}}{p_1} + \frac{p_{n-2}}{p_{n-1}}, \ldots, |a_{n-2}|\frac{p_2}{p_1} + \frac{p_2}{p_3},\ |a_{n-1}| + \frac{p_1}{p_2}\right\}, \qquad (5.6.44)$$

which holds for any positive parameters $p_1, p_2, \ldots, p_n$.

34. If all the coefficients a_k in (5.6.38) are nonzero, choose $p_k \equiv p_1/|a_{n-k+1}|$, $k = 2, 3, \ldots, n$ and deduce *Kojima's bound* on the zeroes $\tilde z$ of $p(z)$ from (5.6.44):

$$|\tilde z| \le \max\left\{|a_0|,\ 2\left|\frac{a_1}{a_2}\right|,\ 2\left|\frac{a_2}{a_3}\right|, \ldots, 2\left|\frac{a_{n-1}}{a_n}\right|\right\} \qquad (5.6.45)$$

35. Now choose $p_k \equiv r^k$, $k = 1, 2, \ldots, n$ for some $r > 0$ and show that (5.6.44) implies the bound

$$|\tilde z| \le \max\{|a_0|r^{n-1},\ |a_1|r^{n-2} + r^{-1},\ |a_2|r^{n-3} + r^{-1}, \ldots, |a_{n-2}|r + r^{-1},\ |a_{n-1}| + r^{-1}\}$$
$$\le \frac{1}{r} + \max_{0 \le k \le n-1} \{|a_k| r^{n-k-1}\} \quad \text{for any } r > 0 \qquad (5.6.46)$$

36. If $A \in M_n$, show that the Hermitian matrix

$$\hat A = \begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} \in M_{2n}$$

has the same spectral norm ($\|\cdot\|_2$) as A. *Hint:* Recall that $\|\hat{A}\|_2 = \rho(\hat{A}^*\hat{A})^{1/2}$ in general.

37. If $A, B \in M_n$, if A is nonsingular, if B is singular, and if $\|\cdot\|$ is any matrix norm, show that $\|A-B\| \geq 1/\|A^{-1}\|$. *Hint:* $B = A-(A-B) = A[I-A^{-1}(A-B)]$ is singular, so $\|A^{-1}(A-B)\| \geq 1$. What does this mean geometrically in M_n? How closely can a nonsingular matrix be approximated by a singular matrix? See (7.4.1) for more information about this question.

Further Readings. The bounds in the table in Problem 23 come from B. J. Stone, "Best Possible Ratios of Certain Matrix Norms," *Numerische Math.* 4 (1962), 114–116, which also contains some additional bounds and references. For additional references and more discussion of the use of matrix norms to locate zeroes of polynomials (Problems 27–35), see M. Fujii and F. Kubo, "Operator Norms as Bounds for Roots of Algebraic Equations," *Proc. Japan Acad.* 49 (1973), 805–808. A more general discussion of the problem of determining bounds between induced norms [Theorem (5.6.18)] is in H. Schneider and W. G. Strang, "Comparison Theorems for Supremum Norms," *Numerische Math.* 4 (1962), 15–20. There is a discussion of minimal matrix norms in [Wie].

5.7 Vector norms on matrices

Although all the axioms for a vector norm are necessary for a useful notion of "size" for matrices, for some important applications the submultiplicativity axiom (4) for a matrix norm is not necessary. For example, the very useful limit (5.6.14) is actually true for a class of functions even more general than vector norms, not just for matrix norms. For this reason, we focus here on vector norms on matrices, that is, vector norms (which may not be submultiplicative) on the vector space M_n. Such norms are often called *generalized matrix norms.* We shall denote a generic vector norm on M_n by $\|\cdot\|$ or by $G(\cdot)$ and we begin with some examples of vector norms on M_n that may or may not be matrix norms.

Example 1. If $G(\cdot)$ is a vector norm on M_n, and if $T, S \in M_n$ are nonsingular, then

$$G_{T,S}(A) \equiv G(TAS), \qquad A \in M_n \tag{5.7.1}$$

is a vector norm on M_n. Even if $G(\cdot)$ is a matrix norm, $G_{T,S}(\cdot)$ need not be submultiplicative.

Exercise. Show that $G_{T,S}(\cdot)$ in (5.7.1) is always a vector norm on M_n.

Exercise. Let $S = T = \frac{1}{2}I$, let $G(\cdot) = \|\cdot\|_\infty$, and show that $G_{T,S}(\cdot)$ is not a matrix norm.

Exercise. Show that if $G(\cdot)$ is a matrix norm and $T = S^{-1}$, then $G_{T,S}(\cdot)$ is a matrix norm.

Example 2. The *Hadamard product* of two matrices $A = [a_{ij}]$ and $B = [b_{ij}]$ of the same size is just their element-wise product $A \circ B \equiv [a_{ij}b_{ij}]$. If $H \in M_n$ is a given matrix with no zero entries and if $G(\cdot)$ is any vector norm on M_n, then

$$G_H(A) \equiv G(H \circ A) \tag{5.7.2}$$

is a vector norm on M_n. Even if $G(\cdot)$ is a matrix norm, $G_H(\cdot)$ need not be submultiplicative.

Exercise. Show that $G_H(\cdot)$ in (5.7.2) is always a vector norm.

Exercise. Show that $G_H(\cdot)$ in (5.7.2) may or may not be a matrix norm, depending on the choice of H. Consider the matrix norm $G(\cdot) = \|\cdot\|_1$, and the Hadamard multiplier matrices

$$H_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \quad \text{or} \quad H_2 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \tag{5.7.3}$$

You may wish to consider

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \qquad \text{and} \quad AB \tag{5.7.4}$$

Notice that $G_{H_1}(C) \leq G_{H_2}(C)$ for all $C \in M_2$.

Example 3. The function

$$G\begin{bmatrix} a & b \\ c & d \end{bmatrix} \equiv \tfrac{1}{2}[|a+d| + |a-d| + |b| + |c|] \tag{5.7.5}$$

is a vector norm on M_2.

Exercise. Show that $G(\cdot)$ in (5.7.5) is a vector norm but not a matrix norm. You may wish to consider the matrices (5.7.4).

Example 4. If $A \in M_n$ is a given matrix, the set $F(A) \equiv \{x^*Ax : x \in \mathbf{C}^n$ and $x^*x = 1\}$ is called the *field of values* or *numerical range* of A, and the function

$$r(A) = \max_{x^*x \leq 1} |x^*Ax| = \max_{z \in F(A)} |z| \tag{5.7.6}$$

is called the *numerical radius* of A.

Exercise. Show that the numerical radius $r(A)$ is a vector norm on M_n. *Hint:* The positivity axiom (1a) is the only hard part; see Section (4.1), Problem 6. The numerical radius is not a matrix norm, however. See Problem 10.

Example 5. The l_∞ vector norm on M_n is

$$\|A\|_\infty \equiv \max_{1\le i,j\le n} |a_{ij}| \tag{5.7.7}$$

We saw in Section (5.6) that $\|\cdot\|_\infty$ is a vector norm on M_n but not a matrix norm, but $n\|\cdot\|_\infty$ is a matrix norm.

The preceding examples demonstrate amply that there are indeed many vector norms on M_n that are not matrix norms. Some of these norms, moreover, share some of the properties of matrix norms that follow from submultiplicativity, and some do not. But each vector norm on M_n is equivalent to any matrix norm (in the sense that they have the same convergent sequences); in fact, a somewhat more general result follows immediately from Theorem (5.4.4).

5.7.8 **Theorem.** Let f be a pre-norm on M_n, that is, a real-valued function on M_n that is positive, homogeneous, and continuous, and let $\|\cdot\|$ be a given matrix norm on M_n. Then there exist finite positive constants C_m and C_M such that

$$C_m\|A\| \le f(A) \le C_M\|A\| \tag{5.7.9}$$

for all $A\in M_n$. In particular, these inequalities hold whenever $f(\cdot)$ is a vector norm on M_n.

The equivalence (5.7.9) is often useful in extending facts about matrix norms to vector norms on matrices, or, more generally, to vector prenorms on matrices. For example, the limit (5.6.14) extends in this manner.

5.7.10 **Corollary.** If f is a pre-norm on M_n, then $\lim_{k\to\infty}[f(A^k)]^{1/k}$ exists for all $A\in M_n$ and

$$\lim_{k\to\infty}[f(A^k)]^{1/k} = \rho(A)$$

for all $A\in M_n$. In particular, this limit holds whenever $f(\cdot)$ is a vector norm on M_n.

Proof: Let $\|\cdot\|$ be a matrix norm on M_n and consider the inequality

$$C_m\|A^k\| \le f(A^k) \le C_M\|A^k\|$$

which implies

$$C_m^{1/k}\|A^k\|^{1/k} \le [f(A^k)]^{1/k} \le C_M^{1/k}\|A^k\|^{1/k}$$

for all $k=1,2,3,\ldots$. But $C_m^{1/k}\to 1$, $C_M^{1/k}\to 1$, and $\|A^k\|^{1/k}\to\rho(A)$ as $k\to\infty$, so we conclude that $\lim_{k\to\infty}[f(A^k)]^{1/k}$ exists and has the asserted value. □

There is a second sense in which any vector norm on M_n is equivalent to a matrix norm, and it is illustrated in Example 5, above. The vector norm $\|\cdot\|_\infty$ can be modified by the constant factor n to make it into a matrix norm. This is no accident: Every vector norm can be so modified.

5.7.11 **Theorem.** For each vector norm $G(\cdot)$ on M_n, there is a finite positive constant $c(G)$ such that $c(G)G(\cdot)$ is a matrix norm on M_n. If $\|\cdot\|$ is a matrix norm on M_n, and if

$$C_m\|A\| \le G(A) \le C_M\|A\| \quad \text{for all} \quad A\in M_n \tag{5.7.11a}$$

then

$$c(G) \le \frac{C_M}{C_m^2}$$

Moreover, there exists a matrix norm for which this upper bound on $c(G)$ is sharp, so

$$c(G) = \min\left\{\frac{C_M}{C_m^2} : \|\cdot\| \text{ is a matrix norm and (5.7.11a) holds}\right\}$$

Proof: For any $c>0$, the function $\|\cdot\| \equiv cG(\cdot)$ satisfies all the axioms for a matrix norm except perhaps the submultiplicative axiom. However, one deduces easily from continuity of $G(\cdot)$ and compactness of the unit ball of $G(\cdot)$ that

$$c(G) \equiv \max_{A\ne 0\ne B} \frac{G(AB)}{G(A)G(B)} = \max_{G(A)=1=G(B)} G(AB)$$

is finite and positive. Then

$$G(AB) \le c(G)G(A)G(B) \quad \text{and} \quad c(G)G(AB) \le c(G)G(A)c(G)G(B)$$

for all $A,B\in M_n$. Suppose $\|\cdot\|$ is a matrix norm on M_n and assume the given inequalities between $G(\cdot)$ and $\|\cdot\|$. Then

$$G(AB) \le C_M\|AB\| \le C_M\|A\|\|B\| \le \frac{C_M}{C_m^2}G(A)G(B)$$

and hence

$$c(G) \le \frac{C_M}{C_m^2}$$

If we take for the matrix norm the particular choice $\|\cdot\| = c(G)G(\cdot)$, then $C_M = c(G)$ and $C_m = 1/c(G)$, so $C_M/C_m^2 = c(G)$. □

Exercise. Show that if $k \ge c(G)$, then $kG(\cdot)$ is a matrix norm. In particular, show that $C_M G(\cdot)/C_m^2$ is always a matrix norm.

Exercise. Deduce the result for vector norms in (5.7.10) directly from (5.7.11).

One of the consequences of submultiplicativity of matrix norms is the fact that associated with every matrix norm on M_n there is some compatible vector norm on $\mathbf{C}^n$. It is a consequence of this that $\|A\| \ge \rho(A)$ for every matrix norm $\|\cdot\|$; a vector norm on M_n that satisfies this inequality for all $A \in M_n$ is said to be *spectrally dominant*. It is interesting to observe that some vector norms on M_n have compatible vector norms on $\mathbf{C}^n$ and some do not. Among those that do not, some are spectrally dominant and some are not. And a vector norm on M_n can have a compatible vector norm on $\mathbf{C}^n$ without being a matrix norm.

5.7.12 Definition. The vector norm $\|\cdot\|$ on $\mathbf{C}^n$ is said to be *compatible* with the vector norm $G(\cdot)$ on M_n if

$$\|Ax\| \le G(A)\|x\|$$

for all $x \in \mathbf{C}^n$ and all $A \in M_n$. The term *consistent* is sometimes used, and the vector norm $\|\cdot\|$ is sometimes said to be *subordinate* to the generalized matrix norm $G(\cdot)$. These ideas have already been touched upon in the previous section [e.g., (5.6.27), (5.6.30)]. We emphasize the observations relevant here.

5.7.13 Theorem. If $\|\cdot\|$ is a matrix norm on M_n, then there is some vector norm on $\mathbf{C}^n$ that is compatible with it.

Proof: If one defines $\|x\| \equiv \|[x\ 0\ 0 \ldots 0]\|$, then $\|Ax\| = \|[Ax\ 0 \ldots 0]\| = \|A[x\ 0 \ldots 0]\| \le \|A\|\,\|[x\,0\ldots 0]\| = \|A\|\,\|x\|$. □

We already know the converse. Theorem (5.6.2) says that if $\|\cdot\|$ is a given vector norm on $\mathbf{C}^n$, then there is a matrix norm [the induced norm (5.6.1)] that is compatible with it.

Exercise. Show that the compatible vector norm on $\mathbf{C}^n$ guaranteed by (5.7.13) need not be unique. Indeed, $\|x\| \equiv \|[x\ x \ldots x]\|$ also works, as does $\|x\| \equiv \|x^*y\|$ for any nonzero vector $y \in \mathbf{C}^n$.

5.7.14 Theorem. Let $G(\cdot)$ be a vector norm on M_n that has a compatible vector norm $\|\cdot\|$ on $\mathbf{C}^n$. Then $G(A) \ge \rho(A)$ for all $A \in M_n$. More generally,

$$G(A_1)G(A_2)\cdots G(A_k) \ge \rho(A_1A_2\cdots A_k) \tag{5.7.15}$$

for all $A_1, A_2, \ldots, A_k \in M_n$ and all $k = 1, 2, \ldots$.

Proof: Suppose $k=2$ and let $x \in \mathbf{C}^n$ be a nonzero vector such that $A_1A_2x = \lambda x$ with $|\lambda| = \rho(A_1A_2)$. Then

$$\rho(A_1A_2)\|x\| = \|\lambda x\| = \|A_1A_2x\| = \|A_1(A_2x)\|$$
$$\le G(A_1)\|A_2x\| \le G(A_1)G(A_2)\|x\|$$

Since $\|x\| \ne 0$, we conclude that $\rho(A_1A_2) \le G(A_1)G(A_2)$. The general case follows in the same way by induction. □

When does a given vector norm on M_n have a compatible vector norm on $\mathbf{C}^n$? The condition (5.7.15) is necessary; to show that it is also sufficient we need a technical lemma.

5.7.16 Lemma. Let $G(\cdot)$ be a vector norm on M_n that satisfies (5.7.15), and let $\|\cdot\|_2$ denote the spectral norm on M_n. Then there is a finite positive constant $c = c(G)$ such that

$$G(A_1)G(A_2)\cdots G(A_k) \ge c\|A_1A_2\cdots A_k\|_2$$

for all $A_1, A_2, \ldots, A_k \in M_n$ and all $k = 1, 2, \ldots$.

Proof: By Corollary (5.4.5) there is a finite positive constant $b = b(G)$ such that $\|A\|_2 \ge bG(A)$ for all $A \in M_n$. Let k be a given positive integer and let $A_1, A_2, \ldots, A_k \in M_n$ be given. By the singular value decomposition theorem (7.3.5) there are unitary matrices V and W and a diagonal matrix $\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ with all $\sigma_i \ge 0$ such that $A_1A_2\cdots A_k = V\Sigma W^*$ and $\rho(\Sigma) = \max\{\sigma_1, \sigma_2, \ldots, \sigma_n\} = \|A_1A_2\cdots A_k\|_2$. By (5.7.15) we have

$$G(V^*)G(A_1)G(A_2)\cdots G(A_k)G(W) \ge \rho(V^*A_1A_2\cdots A_kW)$$
$$= \rho(\Sigma)$$
$$= \|\Sigma\|_2$$
$$= \|V^*A_1A_2\cdots A_kW\|_2$$
$$= \|A_1A_2\cdots A_k\|_2$$

The latter equality is because the spectral norm is unitarily invariant. We conclude that

$$G(A_1)G(A_2)\cdots G(A_k)\ge \frac{1}{G(V^*)G(W)}\|A_1A_2\cdots A_k\|_2$$
$$\ge \frac{b^2}{\|V^*\|_2\|W\|_2}\|A_1A_2\cdots A_k\|_2$$
$$= b^2\|A_1A_2\cdots A_k\|_2$$

If we take $c\equiv b^2$, the assertion of the lemma is proved. □

5.7.17 **Theorem.** Let $G(\cdot)$ be a vector norm on M_n. There is a vector norm $\|\cdot\|$ on $\mathbf{C}^n$ such that

$$\|Ax\|\le G(A)\|x\| \quad \text{for all} \quad x\in\mathbf{C}^n \quad \text{and all} \quad A\in M_n$$

if and only if

$$G(A_1)G(A_2)\cdots G(A_k)\ge \rho(A_1A_2\cdots A_k)$$

for all $A_1, A_2, \ldots, A_k\in M_n$ and all $k=1,2,\ldots$.

Proof: Necessity has already been proved in Theorem (5.7.14). For sufficiency, we shall show that there is a matrix norm $|||\cdot|||$ on M_n such that $G(A)\ge |||A|||$ for all $A\in M_n$. Let $\|\cdot\|$ be a vector norm on $\mathbf{C}^n$ that is compatible with $|||\cdot|||$ [guaranteed to exist by Theorem (5.7.13)] and let $x\in\mathbf{C}^n$ and $A\in M_n$ be given. Then $\|Ax\|\le |||A|||\,\|x\|\le G(A)\|x\|$, so we are done if we can construct a matrix norm that is dominated by $G(\cdot)$.

For a given matrix $A\in M_n$, there are myriad ways to represent A as a product of matrices or as a sum of products of matrices. We define

$$|||A|||\equiv \inf\left\{\sum_i G(A_{i1})\cdots G(A_{ik_i}) : \sum_i A_{i1}\cdots A_{ik_i}=A \text{ and all } A_{ik_j}\in M_n\right\}$$

If $\sum A_{i1}\cdots A_{ik_i}=A$, then by Lemma (5.7.16) and the triangle inequality for the spectral norm we have

$$\sum_i G(A_{i1})\cdots G(A_{ik_i})\ge \sum_i c\|A_{i1}\cdots A_{ik_i}\|_2$$
$$\ge c\left\|\sum_i A_{i1}\cdots A_{ik_i}\right\|_2 = c\|A\|_2$$

From this inequality it follows that the constructed function $|||\cdot|||$ is positive. Homogeneity of $|||\cdot|||$ follows immediately from homogeneity of $G(\cdot)$. The triangle inequality and submultiplicativity for $|||\cdot|||$ follow from its definition as the infimum of sums of products. □

Exercise. Provide the details for the argument that the function $|||\cdot|||$ constructed in the previous theorem obeys the triangle inequality and is submultiplicative. *Hint:* If $C=A+B$ or $C=AB$, then every representation of A and B (separately) as a sum of products yields a representation of C as a sum of products, but not all such representations of C arise in this way.

Exercise. Consider the vector norm (5.7.5) on M_2. If it had a compatible vector norm $\|\cdot\|$ on $\mathbf{C}^2$, then show that

$$\left\|\begin{bmatrix}1\\0\end{bmatrix}\right\|=\left\|\begin{bmatrix}0&1\\0&0\end{bmatrix}\begin{bmatrix}0\\1\end{bmatrix}\right\|\le G\begin{bmatrix}0&1\\0&0\end{bmatrix}\left\|\begin{bmatrix}0\\1\end{bmatrix}\right\|$$

and

$$\left\|\begin{bmatrix}0\\1\end{bmatrix}\right\|=\left\|\begin{bmatrix}0&0\\1&0\end{bmatrix}\begin{bmatrix}1\\0\end{bmatrix}\right\|\le G\begin{bmatrix}0&0\\1&0\end{bmatrix}\left\|\begin{bmatrix}1\\0\end{bmatrix}\right\|$$

which implies that

$$\left\|\begin{bmatrix}1\\0\end{bmatrix}\right\|\le G\begin{bmatrix}0&1\\0&0\end{bmatrix}G\begin{bmatrix}0&0\\1&0\end{bmatrix}\left\|\begin{bmatrix}1\\0\end{bmatrix}\right\|$$

and hence

$$1\le G\begin{bmatrix}0&1\\0&0\end{bmatrix}G\begin{bmatrix}0&0\\1&0\end{bmatrix}$$

Show that this is not correct and conclude that this vector norm $G(\cdot)$ on M_2 cannot have a compatible vector norm on $\mathbf{C}^2$.

Exercise (continued). Even though the vector norm (5.7.5) on M_2 does not have any compatible vector norm on $\mathbf{C}^2$, show directly that it is spectrally dominant. Discuss in light of Theorem (5.7.17).

We now know useful necessary and sufficient conditions for a vector norm on M_n to have a compatible vector norm on $\mathbf{C}^n$. We also know that whenever one has a vector norm on $\mathbf{C}^n$, then the induced matrix norm (5.6.1) is a submultiplicative vector norm on M_n that is compatible with it. When does a vector norm on $\mathbf{C}^n$ have a compatible vector norm on M_n that is not submultiplicative? Always.

5.7.18 **Theorem.** Let $\|\cdot\|$ be a given vector norm on $\mathbf{C}^n$. Then there is a vector norm $G(\cdot)$ on M_n that is *not* a matrix norm and is such that

$$\|Ax\| \le G(A)\|x\|$$

for all $x \in \mathbf{C}^n$ and all $A \in M_n$.

Proof: Let $P \in M_n$ be any permutation matrix with a zero main diagonal – for example, $P=[p_{ij}]$ with $p_{ij}=1$ if $j=i+1$ or if $i=n$ and $j=1$; otherwise, $p_{ij}=0$. Let $\|\!|\cdot|\!\|$ denote the matrix norm on M_n which is induced [via (5.6.1)] by the vector norm $\|\cdot\|$. Define $G(\cdot)$ on M_n by

$$G(A) \equiv \|\!|A|\!\| + \|\!|P|\!\|\,\|\!|P^T|\!\| \max_{1\le i\le n} |a_{ii}|$$

Clearly, $G(\cdot)$ is a vector norm on M_n, $G(A) \ge \|\!|A|\!\|$ for all $A \in M_n$, and

$$\|Ax\| \le \|\!|A|\!\|\,\|x\| \le G(A)\|x\|$$

for all $A \in M_n$ and all $x \in \mathbf{C}^n$. But

$$G(PP^T) = G(I) = \|\!|I|\!\| + \|\!|P|\!\|\,\|\!|P^T|\!\| = 1 + \|\!|P|\!\|\,\|\!|P^T|\!\|$$

$$G(P) = \|\!|P|\!\|$$

$$G(P^T) = \|\!|P^T|\!\|$$

$$G(PP^T) > G(P)G(P^T)$$

Thus, the norm $G(\cdot)$ on M_n is compatible with the given vector norm $\|\cdot\|$ on $\mathbf{C}^n$, but it is not submultiplicative. □

Exercise. Let $A=[a_{ij}] \in M_n$ and consider the following modification of the maximum row sum matrix norm:

$$\|\!|A|\!\| \equiv \|\!|A + \mathrm{diag}(a_{11}, a_{22}, \ldots, a_{nn})|\!\|_\infty$$

Show that this is a Hadamard product norm of the form (5.7.2) and hence is a vector norm on M_n. Show that this norm is compatible with the vector norm $\|\cdot\|_\infty$ on $\mathbf{C}^n$. Now compute

$$\left\|\!\left|\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\right|\!\right\| \quad \text{and} \quad \left\|\!\left|\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}^2\right|\!\right\|$$

and show that this norm is not submultiplicative.

Problems

1. Let $G(\cdot)$ be a vector norm on M_n and let $y \in \mathbf{C}^n$ be a given nonzero vector. Show that the function

$$\|x\| \equiv G(xy^*)$$

is a vector norm on $\mathbf{C}^n$. What is this when

$$y = [1, 1, \ldots, 1]^T \quad \text{or} \quad y = [1, 0, 0, \ldots, 0]^T$$

2. Let $\|\!|\cdot|\!\|$ be any vector norm on M_n, let $A \in M_n$ be given, and let $\epsilon > 0$ be given. Show that there exists some $K = K(\epsilon, A) > 0$ such that

$$[\rho(A) - \epsilon]^k \le \|\!|A^k|\!\| \le [\rho(A) + \epsilon]^k$$

for all $k > K$.

3. Let $\|\!|\cdot|\!\|$ be any vector norm on M_n, and let $A \in M_n$ be given. (a) Use Problem 2 to show that if $\rho(A) < 1$, then $\|\!|A^k|\!\| \to 0$ as $k \to \infty$. At what rate? (b) Conversely, if $\|\!|A^k|\!\| \to 0$ as $k \to \infty$, show that $\rho(A) < 1$. *Hint:* Consider $\|\!|A^k[x \ldots x]|\!\|$ if $Ax = \lambda x$ and $x \ne 0$. (c) What can you say about convergence of power series of matrices using vector norms?

4. Let $G(\cdot)$ be a given vector norm on M_n, and define the function $G': M_n \to \mathbf{R}$ by

$$G'(B) \equiv \max_{G(A)=1} G(BA)$$

Show that $G'(\cdot)$ is always a matrix norm on M_n. Show that $G'(I) = 1$ always. If $G(I) = 1$, show that $G'(B) \ge G(B)$ for all $B \in M_n$.

The next four problems extend Problem 4.

5. If $G(\cdot)$ is a matrix norm on M_n, show that $G'(B) \le G(B)$ for all $B \in M_n$ and if $G(I) = 1$, then $G'(\cdot) = G(\cdot)$.

6. Show that $G''(\cdot) = G'(\cdot)$ always.

7. If $G(\cdot)$ is a vector norm on M_n such that $G(I) = 1$, show that $G(\cdot)$ is a matrix norm if and only if $G'(B) \le G(B)$ for all $B \in M_n$.

8. Show that one could reverse the order in which A and B appear in the definition of $G'(\cdot)$ in Problem 4 and thereby obtain another matrix norm. Show with an example that this other norm need not be equal to $G'(\cdot)$.

9. Show that the set of all vector seminorms on $\mathbf{C}^n$ that are compatible with a given vector norm on M_n is a convex set – it is in fact a cone.

10. Show that the numerical radius $r(\cdot)$ is not a matrix norm on M_n by considering the matrices (5.7.4) and comparing $r(AB)$ with $r(A)r(B)$.

11. The inequality $\|\!|A|\!\| \ge \rho(A)$ in Theorem (5.6.9) follows from the submultiplicativity axiom (4) for a matrix norm $\|\!|\cdot|\!\|$. But it is possible for a vector norm on M_n to satisfy this inequality (i.e., to be spectrally dominant) without being a matrix norm. Show that $r(A) \ge \rho(A)$ for every $A \in M_n$. Show more generally that $\sigma(A) \subset \{x^*Ax : x^*x = 1\}$.

12. Show that the vector norm $\|\cdot\|_\infty$ on M_n cannot have any compatible vector norm on $\mathbf{C}^n$. *Hint:* Consider $\|J_n\|_\infty$ and $\rho(J_n)$. However, show

that $n\|\cdot\|_\infty$ is a matrix norm on M_n, and hence it has a compatible vector norm on $\mathbf{C}^n$.

13. For $A=[a_{ij}]\in M_{m,n}$, denote the transpose of the ith row of A by $r_i(A)=[a_{i1},a_{i2},\ldots,a_{in}]^T$ and the jth column of A by $c_j(A)=[a_{1j},a_{2j},\ldots,a_{mj}]^T$, and suppose that $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ are vector norms on $\mathbf{C}^n$ and $\mathbf{C}^m$, respectively. Then define $G_{\beta,\alpha}: M_{m,n}\to\mathbf{R}$ by

$$G_{\beta,\alpha}(A)\equiv\|[\|r_1(A)\|_\alpha,\|r_2(A)\|_\alpha,\ldots,\|r_m(A)\|_\alpha]^T\|_\beta$$

Similarly, define $G^{\alpha,\beta}: M_{m,n}\to\mathbf{R}$ by

$$G^{\alpha,\beta}(A)\equiv\|[\|c_1(A)\|_\beta,\|c_2(A)\|_\beta,\ldots,\|c_n(A)\|_\beta]^T\|_\alpha$$

Show that $G_{\beta,\alpha}(\cdot)$ and $G^{\alpha,\beta}(\cdot)$ are each vector norms on $M_{m,n}$, but that $G^{\alpha,\beta}(\cdot)$ is not necessarily the same as $G_{\alpha,\beta}(\cdot)$. Note that these are natural ways to define vector norms on the space of rectangular matrices.

14. Compare $G_{\beta,\alpha}(\cdot)$ of Problem 13 to the norms $\|\cdot\|_{\alpha,\beta}$ defined in Problem 4 of Section (5.6), and show by example that even when $m=n$ (and even when $\|\cdot\|_\alpha=\|\cdot\|_\beta$), $G_{\beta,\alpha}(\cdot)$ need not be a matrix norm on M_n.

15. In Problem 13, when $\|\cdot\|_\alpha=\|\cdot\|_2=\|\cdot\|_\beta$, what norm is $G_{\beta,\alpha}(\cdot)$? How about $G^{\alpha,\beta}(\cdot)$?

16. In Problem 13, when $\|\cdot\|_\alpha=\|\cdot\|_1$ and $\|\cdot\|_\beta=\|\cdot\|_\infty$, what norm is $G_{\beta,\alpha}(\cdot)$? How about $G^{\beta,\alpha}(\cdot)$? What about $G_{\alpha,\beta}(\cdot)$ and $G^{\alpha,\beta}(\cdot)$?

17. If $G(\cdot)$ is a vector norm on M_n, the *spectral characteristic* of $G(\cdot)$ is defined as

$$m(G)\equiv\max_{G(A)\le 1}\rho(A)$$

Show that $G(\cdot)$ is spectrally dominant if and only if $m(G)\le 1$, and show that any vector norm on M_n may be converted into a spectrally dominant norm via multiplication by a constant – with the minimum constant necessary being $m(G)$. A norm $G(\cdot)$ on M_n is called *minimally spectrally dominant* if $m(G)=1$.

18. Show that any induced matrix norm is minimally spectrally dominant, as defined in Problem 17. Show that there are norms that are minimally spectrally dominant but are not induced. Show that the numerical radius $r(A)$ is minimally spectrally dominant.

19. Show that the spectral characteristic is a convex function on the cone of vector norms on M_n and therefore that the set of all spectrally dominant vector norms on M_n is convex.

20. Show that a vector norm $G(\cdot)$ on M_n is spectrally dominant if and only if for each $A\in M_n$ there is a constant γ_A [depending only upon $G(\cdot)$ and A] such that for all integers $k>0$,

$$G(A^k)\le\gamma_A G(A)^k$$

21. (a) Show that the numerical radius $r(\cdot)$ satisfies $r(A)=\rho(A)=\|\!|A|\!\|_2$ whenever A is normal, but that $r(A)\le\|\!|A|\!\|_2$ in general. Give an example of a matrix $A\in M_n$ such that $r(A)<\|\!|A|\!\|_2$. *Hint:* Show that $r(U^*AU)=r(A)$ whenever $U\in M_n$ is unitary, and use the fact that A is unitarily diagonalizable. Then observe that

$$r(A)=\max_{\|x\|_2=1}|x^*Ax|\le\max_{\|x\|_2=1}\|Ax\|_2\|x\|_2=\|\!|A|\!\|_2$$

in general. (b) Show that $r(A)=r(A^*)$ for all $A\in M_n$. (c) Show that $\|\!|A|\!\|_2\le 2r(A)$ for all $A\in M_n$ as follows: Write

$$A=(A+A^*)/2+(A-A^*)/2\equiv A_1+A_2$$

and observe that A_1 and A_2 are normal. Now show that

$$\|\!|A|\!\|_2\le\|\!|A_1|\!\|_2+\|\!|A_2|\!\|_2=r(A_1)+r(A_2)\le r(A)+r(A^*)=2r(A)$$

(d) Show that the bounds

$$\tfrac{1}{2}\|\!|A|\!\|_2\le r(A)\le\|\!|A|\!\|_2 \qquad \text{(i)}$$

proved in (a) and (c) are sharp by considering suitable n-by-n versions of the matrices $\begin{bmatrix}1&0\\0&0\end{bmatrix}$ and $\begin{bmatrix}0&1\\0&0\end{bmatrix}$.

22. Use the inequalities in (d) of Problem 21 and the bounds on $c(r)$ given in Theorem (5.7.11) to show that the function $4r(\cdot)$ is a matrix norm on M_n. Show that $c(r)=4$ by considering $A=\begin{bmatrix}0&1\\0&0\end{bmatrix}$, A^*, and AA^*.

23. Deduce from (i) in Problem 21(d) and the inequality

$$\frac{1}{\sqrt{n}}\|A\|_2\le\|\!|A|\!\|_2\le\|A\|_2 \qquad \text{(ii)}$$

proved in Problem 23, Section (5.6), that

$$\frac{1}{2\sqrt{n}}\|A\|_2\le r(A)\le\|A\|_2 \qquad \text{(iii)}$$

for all $A\in M_n$ and show that the upper bound is sharp. Verify that $A=\begin{bmatrix}0&1\\0&0\end{bmatrix}$ and $A=I$ are examples of equality in the lower bounds of (i) and (ii), respectively, and that $A=\begin{bmatrix}1&0\\0&0\end{bmatrix}$ is an example of equality in the upper bounds of (i) and (ii). Explain why the upper bound in (iii) must,

therefore, be sharp, and give an example of a case of equality. The lower bound in (iii), however, is not sharp. Why is there a finite maximal positive constant c_n such that $c_n\|A\|_2 \le r(A)$ for all $A \in M_n$? Actually, $c_n = (2n)^{-1/2}$ for even n and $c_n = (2n-1)^{-1/2}$ for odd n. For even n the cases of equality are the matrices unitarily similar to a direct sum of matrices of the form $r(A)\left[\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right]$; an additional single 1-by-1 direct summand $[\alpha]$, $|\alpha| = r(A)$, must be included when n is odd.

24. Show that $[A, B] \equiv \operatorname{tr} AB^*$ defines an inner product on M_n and that the l_2 norm on M_n is derived from $[\cdot, \cdot]$; that is, $\|A\|_2 = [A, A]^{1/2}$ for all $A \in M_n$. Show that if $X = xx^*$ is a Hermitian rank 1 matrix, then $\|X\|_2 = \|x\|_2^2$. Show that the field of values of a given matrix $A \in M_n$ is just the set of projections (in the inner product $[\cdot, \cdot]$) of A onto the set of unit norm rank 1 Hermitian matrices, and that $r(A) = \max\{|[A, X]| : X$ is a rank 1 Hermitian matrix and $\|X\|_2 = 1\}$. Use the Cauchy–Schwarz inequality to show that $r(A) \le \|A\|_2$.

25. The numerical radius is related to a natural approximation problem. Let $A \in M_n$ be given, and suppose we wish to approximate A as well as possible in the sense of least squares by a scalar multiple of a Hermitian matrix of rank 1. If we write $X = cxx^*$, $c \in \mathbf{C}$, $\|x\|_2 = 1$, then show that

$$\|A - X\|_2^2 = \|A - cxx^*\|_2^2 \ge \|A\|_2^2 - 2|c[A, xx^*]| + |c|^2$$

is minimized when $c = [A, \tilde{x}\tilde{x}^*]$ and $\tilde{x}$ is a unit vector for which the maximum imum in (5.7.6) is achieved. Conclude that if $\|A - cX\|_2$ is minimum among all scalars c and all rank 1 Hermitian matrices X with $\|X\|_2 = 1$, then $|c| = r(A)$.

26. The preceding two problems suggest a natural generalization of the numerical radius and the field of values. Let $\Phi \subset M_n$ be a nonempty set of matrices such that

(a) If $X \in \Phi$, then $aX \in \Phi$ for all $a \in \mathbf{C}$;
(b) $[A, X] = \operatorname{tr} AX^* = 0$ for all $X \in \Phi$ if and only if $A = 0$; and
(c) Φ is a closed set.

If $A \in M_n$, define

$$\phi(A) \equiv \max_{\substack{X \in \Phi \\ \|X\|_2 \le 1}} |[A, X]| = \max_{\substack{X \in \Phi \\ \|X\|_2 \le 1}} |\operatorname{tr} AX^*|$$

Show that $\phi(\cdot)$ is well defined, is a vector norm on M_n, and satisfies $|\phi(A)| \le \|A\|_2$. Show that for each $A \in M_n$ there is some $X_A \in \Phi$ such that $\|X_A\|_2 = 1$ and $\phi(A) = |[A, X_A]|$.

Consider the problem of approximating a given $A \in M_n$ by matrices from Φ. That is, find some $X \in \Phi$ such that $\|A - X\|_2$ is minimized. Show

that a best approximation is given by the matrix $\phi(A)X_A$, where $\phi(A) = |[A, X_A]|$, and that the error in the approximation of A by any $X \in \Phi$ has the sharp bound $\|A - X\|_2^2 \ge \|A\|_2^2 - |[A, X_A]|^2 \ge 0$.

Show that if Φ is the set of all scalar multiples of rank 1 Hermitian matrices, then $\phi(A) = r(A)$. The case in which Φ is the set of all scalar multiples of unitary matrices is discussed in Example (7.4.6); in this case $\phi(A)$ is the average of the singular values of A. Another interesting case is when Φ is the set of all singular matrices, discussed in Example (7.4.1); in this case $\phi(A)$ is the smallest singular value of A. Other interesting cases are the set of all scalar multiples of positive definite, Hermitian, or normal matrices of a given rank, or the set of all scalar multiples of all matrices which are unitarily similar to a given matrix. In each of these cases, the analog of the field of values is the set $\{[A, X] : X \in \Phi\}$.

27. Even though the numerical radius $r(A)$ is not a matrix norm, it does satisfy the power inequality $r(A^m) \le [r(A)]^m$ for all $m = 1, 2, \ldots$ and all $A \in M_n$. Prove this with the following steps:

(a) Show that it is sufficient to prove that if $r(A) \le 1$, then $r(A^m) \le 1$ for all $m = 1, 2, \ldots$.

Let $m \ge 2$ be a given positive integer, fixed for the rest of the argument, and let $\{w_k\} = \{e^{2\pi i k/m}\}_{k=1}^m$ denote the set of mth roots of unity. Notice that $\{w_k\}$ is a finite multiplicative group and that $\{w_j w_k\}_{k=1}^m = \{w_k\}_{k=1}^m$ for each $j = 1, 2, \ldots, m$.

(b) Observe that

$$1 - z^m = \prod_{k=1}^m (1 - w_k z)$$

and show that

$$p(z) \equiv \frac{1}{m} \sum_{j=1}^m \prod_{\substack{k=1 \\ k \ne j}}^m (1 - w_k z) \equiv 1 \quad \text{for all} \quad z \in \mathbf{C}$$

Hint: Notice that $p(z)$ is a polynomial of degree at most $m - 1$, and that

$$p(z) = \frac{1}{m} \sum_{j=1}^m \frac{1 - z^m}{1 - w_j z}$$

so that $p(z) = p(w_1 z) = \cdots = p(w_m z)$ for all $z \in \mathbf{C}$. Hence $p(z) \equiv$ constant $= p(0) = 1$.

(c) Show that

$$I - A^m = \prod_{k=1}^m (I - w_k A) \quad \text{and} \quad I = \frac{1}{m} \sum_{j=1}^m \prod_{\substack{k=1 \\ k \ne j}}^m (I - w_k A)$$

(d) Let $x \in \mathbf{C}^n$ be any unit vector, $\|x\|_2 = 1$, and let $A \in M_n$. Verify that

$$1-x^*A^m x = x^*(I-A^m)x = (Ix)^*(I-A^m)x$$

$$= \left[\frac{1}{m}\sum_{j=1}^{m}\prod_{\substack{k=1\\k\neq j}}^{m}(I-w_kA)x\right]^*\left[\prod_{k=1}^{m}(I-w_kA)x\right]$$

$$= \frac{1}{m}\sum_{j=1}^{m} z_j^*[(I-w_jA)z_j], \qquad z_j \equiv \prod_{\substack{k=1\\k\neq j}}^{m}(I-w_kA)x$$

$$= \frac{1}{m}\sum_{\substack{j=1\\z_j\neq 0}}^{m}\|z_j\|_2^2\left[1-w_j\left(\frac{z_j}{\|z_j\|_2}\right)^*A\left(\frac{z_j}{\|z_j\|_2}\right)\right]$$

(e) Now replace A by $e^{i\theta}A$ in the identity in (d) to get

$$1-e^{im\theta}x^*A^m x = \frac{1}{m}\sum_{\substack{j=1\\z_j\neq 0}}^{m}\|z_j\|_2^2\left[1-e^{i\theta}w_j\left(\frac{z_j}{\|z_j\|_2}\right)^*A\left(\frac{z_j}{\|z_j\|_2}\right)\right]$$

for any real θ. Now suppose that $r(A)\leq 1$, and show that the real part of the right-hand side of this identity is nonnegative for any $\theta\in\mathbf{R}$, and deduce that the real part of the left-hand side must also be nonnegative for all $\theta\in\mathbf{R}$. Since θ is arbitrary, argue that this implies that $|x^*A^m x|\leq 1$ and hence that $r(A^m)\leq 1$.

28. Even though the numerical radius satisfies the power inequality $r(A^m)\leq r(A)^m$, it is not always true that $r(A^{k+m})\leq r(A^k)r(A^m)$. Verify this by considering $A=J_4(0)$ (the 4-by-4 Jordan block matrix), $k=1$, and $m=2$. *Hint:* Use the arithmetic-geometric mean inequality to show that $r(A^2)=r(A^3)=\frac{1}{2}$ and the Cauchy-Schwarz inequality to show that $r(A)<1$.

29. Is there a sensible notion of "minimal vector norm" on M_n analogous to the concept of minimal matrix norm (5.6.29)?

Further Readings. For more discussion of inequalities involving the numerical radius see M. Goldberg and E. Tadmor, "On the Numerical Radius and Its Applications," *Lin. Alg. Appl.* 42 (1982), 263–284. The proof of the power inequality for the numerical radius in Problem 27 is taken from C. Pearcy, "An Elementary Proof of the Power Inequality for the Numerical Radius," *Michigan Math. J.* 13 (1966), 289–291. Some of the material of this section was developed by C. R. Johnson in "Multiplicativity and Compatibility of Generalized Matrix Norms," *Linear Alg. Appl.* 16 (1977), 25–37, "Locally Compatible Generalized Matrix Norms," *Numer. Math.* 27 (1977), 391–394, and "Power Inequalities and Spectral

Dominance of Generalized Matrix Norms," *Linear Alg. Appl.* 28 (1979), 117–130, where further results may be found.

5.8 Errors in inverses and solutions of linear systems

As an application of matrix and vector norms, we consider the problem of estimating the error made in computing the inverse of a matrix and the solution to a system of linear equations.

If a nonsingular matrix $A\in M_n$ is given, we may imagine that we can compute the inverse matrix A^{-1} exactly, but if the computations are performed on a digital computer with a finite-length machine word, there are inevitable and unavoidable errors of rounding and truncation. Furthermore, even if all the computations could be performed with perfect accuracy, it could be that the elements of the matrix A are the result of some experiment or of some calculation that is subject to errors, and therefore they may not be known with perfect accuracy. How do errors in the computation and errors in the data affect the computed matrix inverse?

It turns out for many common algorithms that round-off errors in the computations can be modeled in the same way as errors in the data. That is, let us suppose $A\in M_n$ is a given nonsingular matrix, and we wish to compute A^{-1}, but what we actually compute is $(A+E)^{-1}$, where $E\in M_n$ is "small" enough so that $A+E$ is invertible. Then the error is $A^{-1}-(A+E)^{-1}=A^{-1}-(I+A^{-1}E)^{-1}A^{-1}$. If $\rho(A^{-1}E)<1$, then $A+E$ will be invertible and we can write $(I+A^{-1}E)^{-1}$ as a power series in $A^{-1}E$. This gives

$$A^{-1}-(A+E)^{-1}=A^{-1}-\sum_{k=0}^{\infty}(-1)^k(A^{-1}E)^kA^{-1}$$

$$=\sum_{k=1}^{\infty}(-1)^{k+1}(A^{-1}E)^kA^{-1}$$

Thus, we have an exact formula for the error

$$A^{-1}-(A+E)^{-1}=\sum_{k=1}^{\infty}(-1)^{k+1}(A^{-1}E)^kA^{-1} \quad \text{if} \quad \rho(A^{-1}E)<1 \qquad (5.8.1)$$

Now suppose that $\|\cdot\|$ is a given matrix norm, and assume that $\|A^{-1}E\|<1$, so that, in particular, $\rho(A^{-1}E)<1$ and (5.8.1) holds. Then

$$\|A^{-1}-(A+E)^{-1}\| = \left\|\sum_{k=1}^{\infty}(-1)^{k+1}(A^{-1}E)^kA^{-1}\right\|$$

$$\leq \sum_{k=1}^{\infty}\|A^{-1}E\|^k\|A^{-1}\| = \frac{\|A^{-1}E\|}{1-\|A^{-1}E\|}\|A^{-1}\|$$

and we conclude that an upper bound on the relative error made in computing the inverse is

$$\frac{\|A^{-1}-(A+E)^{-1}\|}{\|A^{-1}\|} \le \frac{\|A^{-1}E\|}{1-\|A^{-1}E\|} \quad \text{if} \quad \|A^{-1}E\|<1 \tag{5.8.2}$$

If we assume, in addition, that E is "small" enough so that $\|E\| < 1/\|A^{-1}\|$, then $\rho(A^{-1}E) \le \|A^{-1}E\| \le \|A^{-1}\| \, \|E\| < 1$ and we have the estimate

$$\frac{\|A^{-1}-(A+E)^{-1}\|}{\|A^{-1}\|} \le \frac{\|A^{-1}\| \, \|E\|}{1-\|A^{-1}\| \, \|E\|} = \frac{\|A^{-1}\| \, \|A\| (\|E\|/\|A\|)}{1-\|A^{-1}\| \, \|A\| (\|E\|/\|A\|)}$$

The quantity

$$\kappa(A) \equiv \begin{cases} \|A^{-1}\| \, \|A\| & \text{if } A \text{ is nonsingular} \\ \infty & \text{if } A \text{ is singular} \end{cases} \tag{5.8.3}$$

is called the *condition number for matrix inversion with respect to the matrix norm* $\|\cdot\|$. Notice that $\kappa(A) = \|A^{-1}\| \, \|A\| \ge \|A^{-1}A\| = \|I\| \ge 1$ for any matrix norm.

Using this notation, we have the estimate

$$\frac{\|A^{-1}-(A+E)^{-1}\|}{\|A^{-1}\|} \le \frac{\kappa(A)}{1-\kappa(A)(\|E\|/\|A\|)} \frac{\|E\|}{\|A\|} \quad \text{if} \quad \|E\| \, \|A^{-1}\|<1 \tag{5.8.4}$$

which bounds the relative error in the inverse in terms of the relative error in the data. For $\|E\|$ small, the right-hand side is of the order of $\kappa(A)\|E\|/\|A\|$, so we have good reason to believe that the relative error in the inverse is of the same order as the relative error in the data, provided that $\kappa(A)$ is not large. For purposes of inversion, we say that A is *ill conditioned* or *poorly conditioned* (with respect to the matrix norm $\|\cdot\|$) if $\kappa(A)$ is large; if $\kappa(A)$ is small (near 1), we say that A is *well conditioned* (with respect to the matrix norm $\|\cdot\|$); if $\kappa(A)=1$, we say that A is *perfectly conditioned* (with respect to the matrix norm $\|\cdot\|$).

There is an interesting geometric characterization of the condition number in the common case that the norm used is the spectral norm. Let $\theta(A)$ denote the least angle between the vectors Ax and Ay as x and y range over all pairs of orthonormal vectors. Using the spectral norm, $\kappa(A) = \cot[\theta(A)/2]$. Thus, if A is unitary, $\theta(A) = \pi/2$ and $\cot(\pi/4) = 1 = \kappa(A)$. If A is "nearly singular," then there is some orthonormal pair x, y such that Ax is "nearly parallel" to Ay, $\theta(A)$ will be small, and $\kappa(A) = \cot[\theta(A)/2]$ will be large. For more details, see Example (7.4.26).

Exercise. Show that if $A \in M_n$ is invertible, then $\kappa(A) = \kappa(A^{-1})$.

Exercise. Show that if $U, V \in M_n$ are unitary and if the spectral norm (or any other unitarily invariant norm) is used, then $\kappa(A) = \kappa(UA) = \kappa(AV) = \kappa(UAV)$. Thus, unitary transformations of a given matrix do not make it any more ill conditioned than it is already. This observation underlies many stable numerical linear algebra algorithms.

Exercise. Show that $\kappa(AB) \le \kappa(A)\kappa(B)$ always. Is $\kappa(\cdot)$ a matrix or vector norm on M_n?

These same considerations can be used to give a priori bounds on the accuracy of a solution to a system of linear equations. Suppose one wishes to solve

$$Ax = b, \qquad A \in M_n, \qquad b \in \mathbf{C}^n \tag{5.8.5}$$

but because of computational errors or uncertainty in the data one actually solves

$$(A+E)\hat{x} = b, \qquad A, E \in M_n, \qquad b \in \mathbf{C}^n \tag{5.8.6}$$

What can we say about the error $x - \hat{x}$?

If E is "small" enough so that $\rho(A^{-1}E) < 1$, then by (5.8.1) we have

$$x - \hat{x} = A^{-1}b - (A+E)^{-1}b = [A^{-1} - (A+E)^{-1}]b$$

$$= \sum_{k=1}^{\infty} (-1)^{k+1}(A^{-1}E)^k A^{-1}b = \sum_{k=1}^{\infty} (-1)^{k+1}(A^{-1}E)^k x$$

If $\|\cdot\|$ is a matrix norm such that $\|A^{-1}E\| < 1$, and if $\|\cdot\|$ is a compatible vector norm, then an upper bound on the norm of the error is

$$\|x - \hat{x}\| \le \sum_{k=1}^{\infty} \|A^{-1}E\|^k \|x\| = \frac{\|A^{-1}E\|}{1-\|A^{-1}E\|}\|x\|$$

In terms of relative errors, this says that

$$\frac{\|x-\hat{x}\|}{\|x\|} \le \frac{\|A^{-1}E\|}{1-\|A^{-1}E\|} \quad \text{if} \quad \|A^{-1}E\| < 1 \tag{5.8.7}$$

and if $\|\cdot\|$ is a vector norm that is compatible with the matrix norm $\|\cdot\|$. Notice that this is the same as the upper bound (5.8.2) on the relative error of the inverse, and it is independent of the right-hand-side b of the system of linear equations.

In terms of the condition number of A, the same argument used to derive (5.8.4) shows that the relative error in the solution to (5.8.5) has the bound

$$\frac{\|x-\hat{x}\|}{\|x\|} \le \frac{\kappa(A)}{1-\kappa(A)(\|E\|/\|A\|)} \frac{\|E\|}{\|A\|} \quad \text{if} \quad \|A^{-1}\| \, \|E\| < 1 \tag{5.8.8}$$

and if the vector norm $\|\cdot\|$ is compatible with the matrix norm $|||\cdot|||$. Whatever algorithm is used to solve the linear equations (5.8.5), the relative error in the solution has the same bound as the relative error in the inverse of the matrix of coefficients.

It may be that the ideal system of linear equations (5.8.5) is in practice subject to uncertainty about the elements of the right-hand-side b as well as about the elements of the coefficient matrix A. Thus, we may wish to replace (5.8.6) with

$$(A+E)\hat{x}=b+e \tag{5.8.9}$$

where $E\in M_n$ and $e\in \mathbf{C}^n$ are thought of as "small" errors in the data.

Using the same methods, one finds (if $b\neq 0$) that a bound for the relative error between the solutions to (5.8.5) and (5.8.9) is given by

$$\frac{\|x-\hat{x}\|}{\|x\|}\leq \frac{\kappa(A)}{1-\kappa(A)(|||E|||/|||A|||)}\frac{|||E|||}{|||A|||}+\frac{\kappa(A)}{1-\kappa(A)(|||E|||/|||A|||)}\frac{\|e\|}{\|b\|} \tag{5.8.10}$$

under the same hypotheses as for (5.8.8). Thus, the relative error bound has two terms, one for the relative error in the coefficients A and one for the relative error in the right-hand-side b. The condition number $\kappa(A)$ again plays a crucial role in determining the sensitivity of the bound on the solution error to errors in the data.

All our estimates so far have been a priori bounds on the error; they do not involve the computed solution or any quantity derived from it. Suppose, however, that some computed "solution" $\hat{x}$ to the system (5.8.5) has been found. It may not be the case that $A\hat{x}=b$ exactly, but from the *residual vector* $r\equiv b-A\hat{x}$ we can obtain an estimate of how close $\hat{x}$ is to the true solution x. Since $A^{-1}r=A^{-1}[b-A\hat{x}]=A^{-1}b-\hat{x}=x-\hat{x}$, we have the straightforward bound $\|x-\hat{x}\|\leq\|A^{-1}r\|$. If $|||\cdot|||$ is a matrix norm that is compatible with the vector norm $\|\cdot\|$, we have $\|b\|=\|Ax\|\leq|||A|||\,\|x\|$, or $1\leq|||A|||\,\|x\|/\|b\|$ if $b\neq 0$, and hence

$$\|x-\hat{x}\|\leq|||A^{-1}|||\,\|r\|\leq\frac{|||A|||\,\|x\|}{\|b\|}|||A^{-1}|||\,\|r\|=|||A|||\,|||A^{-1}|||\frac{\|r\|}{\|b\|}\|x\|$$

Thus, if $b\neq 0$, the relative error between the computed solution $\hat{x}$ (such that $A\hat{x}=b-r$) and the true solution x (such that $Ax=b$) has the bound

$$\frac{\|x-\hat{x}\|}{\|x\|}\leq\kappa(A)\frac{\|r\|}{\|b\|} \tag{5.8.11}$$

where we assume that the matrix norm used to compute the condition number $\kappa(A)$ is compatible with the vector norm $\|\cdot\|$. For a well-conditioned problem, the relative error in the solution is not (much) worse

than the relative size of the residual; for an ill-conditioned problem, even a computed solution that yields a small residual may still be very far from the true solution.

As a final remark about norm estimates of errors, we note that the upper bounds derived in this section are just that – *upper* bounds. The upper bound may be large and the actual error may nevertheless be small. A common characteristic of such bounds is their conservatism: They give bounds on the error that are unduly pessimistic for many problems. However, if a matrix A of moderate size with moderate-size elements has a large condition number, then A^{-1} must have some large entries, and it is well to exercise great caution for the following reason.

If $Ax=b$ and if we set $C=[c_{ij}]\equiv A^{-1}$, then differentiating the identity $x=Cb$ with respect to the entry b_j gives the identities

$$\frac{\partial x_i}{\partial b_j}=c_{ij},\qquad i,j=1,2,\ldots,n \tag{5.8.12}$$

Furthermore, if we consider $C=A^{-1}$ as a function of A, then its entries are just rational functions of the entries of A and hence are differentiable. The identity $CA=I$ means that for all $i,\ q=1,\ldots,n$ we have

$$\sum_{p=1}^{n}c_{ip}a_{pq}=\delta_{iq}$$

and hence

$$\sum_{p=1}^{n}\left[\frac{\partial c_{ip}}{\partial a_{jk}}a_{pq}+\delta_{pq,jk}c_{ip}\right]=\left[\sum_{p=1}^{n}\frac{\partial c_{ip}}{\partial a_{jk}}a_{pq}\right]+\delta_{qk}c_{ij}=0$$

or

$$\sum_{p=1}^{n}\frac{\partial c_{ip}}{\partial a_{jk}}a_{pk}=-\delta_{qk}c_{ij},\qquad i,j,k=1,\ldots,n$$

Now differentiate the identity $x=Cb$ with respect to a_{jk} to obtain

$$\frac{\partial x_i}{\partial a_{jk}}=\sum_{p=1}^{n}\frac{\partial c_{ip}}{\partial a_{jk}}b_p=\sum_{p=1}^{n}\sum_{q=1}^{n}\frac{\partial c_{ip}}{\partial a_{jk}}a_{pq}x_q$$

$$=\sum_{q=1}^{n}\left[\sum_{p=1}^{n}\frac{\partial c_{ip}}{\partial a_{jk}}a_{pq}\right]x_q=\sum_{q=1}^{n}[-\delta_{qk}c_{ij}]x_q=-c_{ij}x_k$$

which is the identity

$$\frac{\partial x_i}{\partial a_{jk}}=-c_{ij}\sum_{p=1}^{n}c_{kp}b_p \tag{5.8.13}$$

Thus, (5.8.12) and (5.8.13) together warn us that if $C=A^{-1}$ has any relatively large entries, then some entry of the solution x may have a

large and unavoidable sensitivity to perturbations in some of the entries of b and A.

Problems

1. Show that the condition number for inversion of a nonsingular normal matrix with respect to the spectral norm is

$$\kappa(A) = \rho(A)\rho(A^{-1}) = |\lambda_{\max}(A)/\lambda_{\min}(A)|$$

2. Compute the eigenvalues and the inverse of the matrix

$$A = \begin{bmatrix} 1 & -1 \\ -1 & 1+\epsilon \end{bmatrix}, \qquad \epsilon > 0$$

Show that, as $\epsilon \to 0$, the ratio of the largest to smallest eigenvalues of A is of the order of ϵ^{-1}. Use Problem 1 to conclude that $\kappa(A) = O(\epsilon^{-1})$ with respect to the spectral norm. Argue that $\kappa(A) = O(\epsilon^{-1})$ with respect to *any* norm, so that A is ill conditioned with respect to inversion. Use the explicit form of A^{-1} to verify that $\kappa(A) = O(\epsilon^{-1})$ for any norm.

3. Compute the eigenvalues and the inverse of the matrix

$$B = \begin{bmatrix} 1 & -1 \\ 1 & -1+\epsilon \end{bmatrix}, \qquad \epsilon > 0$$

Show that, as $\epsilon \to 0$, the ratio of the largest to smallest eigenvalues of B is of the order 1. Show, however, that $\kappa(B) = O(\epsilon^{-1})$ with respect to any matrix norm, and hence B is ill conditioned with respect to inversion as $\epsilon \to 0$. Conclude that the ratio of largest to smallest modulus eigenvalues need not be a condition number for non-normal matrices.

4. The condition number $\kappa(A)$ for inversion depends on the matrix norm used, but show that all condition numbers are equivalent in the sense that if $\kappa_\alpha(A) \equiv \|A^{-1}\|_\alpha \|A\|_\alpha$ and $\kappa_\beta \equiv \|A^{-1}\|_\beta \|A\|_\beta$, then there exist finite positive constants C_m and C_M such that

$$C_m \kappa_\alpha(A) \le \kappa_\beta(A) \le C_M \kappa_\alpha(A) \quad \text{for all} \quad A \in M_n$$

5. Show that every unitary matrix U is perfectly conditioned for inversion $[\kappa(U) = 1]$ with respect to the spectral norm. If the l_2 norm is used, however, the condition number $\kappa(U)$ of every unitary matrix $U \in M_n$ is n.

6. Show that $\kappa(A) \ge |\lambda(A)|_{\max}/|\lambda(A)|_{\min}$ for any nonsingular $A \in M_n$ and any matrix norm (max and min refer in this case to the modulus of the eigenvalue). Thus, if this ratio of eigenvalues is large, the matrix must be ill conditioned for inversion, whether or not it is normal. However, Problem 3 shows that if the matrix is not normal, it can be ill conditioned even if this ratio is not large.

7. Provide the details for the generalization (5.8.10) of the bound (5.8.8), to which the former bound reduces when $e = 0$.

8. Let x be a unit vector in $\mathbf{C}^n$ and let $\lambda > 0$. Show that $A \equiv I + \lambda x x^*$ is a Hermitian matrix with eigenvalues 1 (with multiplicity $n-1$) and $1+\lambda$. Show that $\kappa(A) = 1 + \lambda$ (with respect to the spectral norm), so this gives a simple method to produce an invertible matrix with bounded entries and arbitrarily large condition number. How?

9. Let B be the matrix in Problem 3, and consider the linear equations $Bx = [1, 1]^T$ with exact solution $x = [1, 0]^T$ and a given approximate solution $\hat{x} \equiv [1+\epsilon^{-1/2}, \epsilon^{-1/2}]^T$. Show that the relative error in the residual is $\|r\|/\|b\| = O(\epsilon^{1/2}) \to 0$ as $\epsilon \to 0$, but that the relative error in the solution is $\|x - \hat{x}\|/\|x\| = O(\epsilon^{-1/2}) \to \infty$ as $\epsilon \to 0$. Thus, small residuals can result from approximate solutions that have large errors. Explain in light of the bound (5.8.11).

10. If $\det A$ is small (or large), must $\kappa(A)$ be large? *Hint:* Consider $A = \lambda I \in M_n$.

11. The result of (5.8.4) is, in general, weaker than that of (5.8.2) because the hypothesis $\|A^{-1}\| \, \|E\| < 1$ is more restrictive than having $\|A^{-1}E\| < 1$. Nevertheless, even if the stronger hypothesis is satisfied, (5.8.2) may still give a better upper bound than (5.8.4). Illustrate this with $E = \epsilon A$, $0 < \epsilon < 1$.

12. Find the analogs of the bounds (5.8.7) and (5.8.8) if the equations (5.8.5) and (5.8.6) are replaced by

$$AX = B \quad \text{and} \quad (A+E)\hat{X} = B$$

where $A, E \in M_n$, and $X, B \in M_{n,k}$. Consider the special case $k = n$ and $B = I$; does this help to "explain" why the upper bounds in (5.8.2) and (5.8.7) are the same?

13. The bounds we have derived on the error in the inverse all rely on (5.8.1), which requires that $\rho(A^{-1}E) < 1$. Show that if A and $A+E$ are both invertible, then

$$\|A^{-1} - (A+E)^{-1}\| \le \|A^{-1}\| \, \|(A+E)^{-1}\| \, \|E\|$$

for any matrix norm $\|\cdot\|$ regardless of the size of $A^{-1}E$. *Hint:* Show that $A^{-1} - (A+E)^{-1} = (A+E)^{-1}EA^{-1}$.

14. Perhaps the most commonly cited example of an ill-conditioned matrix is the Hilbert matrix $H_n = [h_{ij}] \in M_n$ defined by $h_{ij} = 1/(i+j-1)$. Show that the condition number of H_n with respect to the spectral norm is given by $|\lambda_{\max}/\lambda_{\min}|$. It is a fact that the condition number of H_n is

asymptotically equal to e^{cn}, where the constant c is approximately 3.5, and it is also a fact that $\rho(H_n)=\pi+O[1/(\log n)]$ as $n\to\infty$. We have $\kappa(H_3)\sim 5\times 10^2$, $\kappa(H_6)\sim 1.5\times 10^7$, and $\kappa(H_8)\sim 1.5\times 10^{10}$. Explain why H_n is so poorly conditioned even though the elements of H_n are all uniformly bounded and $\rho(H_n)$ is not large.

15. If the spectral norm is used, show that $\kappa(A^*A)=\kappa(AA^*)=[\kappa(A)]^2$. Explain why the problem of solving $A^*Ax=y$ may be intrinsically less tractable numerically than the problem of solving $Ax=z$.

16. Let $A\in M_n$ be nonsingular. Use the inequality in Problem 37 of Section (5.6) to show that $\kappa(A)\geq |||A|||/|||A-B|||$ for any singular $B\in M_n$. Here, $|||\cdot|||$ is any matrix norm and $\kappa(\cdot)$ is the associated condition number. This lower bound can be useful in showing that a given matrix A is ill conditioned.

17. Let $A=[a_{ij}]\in M_n$ be an upper triangular matrix with all $a_{ii}\neq 0$. Use Problem 16 to show that the condition number with respect to the maximum row sum norm has the lower bound

$$\kappa(A)\geq\frac{|||A|||_\infty}{\min_{1\leq i\leq n}|a_{ii}|}$$

Further Reading. The problem of finding a priori bounds for errors in solving linear systems of equations has been a central one in numerical linear algebra; see [Ste].

CHAPTER 6

Location and perturbation of eigenvalues

6.0 Introduction

The eigenvalues of a diagonal matrix are very easy to locate, and the eigenvalues of a matrix are continuous functions of the entries, so it is natural to ask whether one can say anything useful about the eigenvalues of a matrix whose off-diagonal elements are "small" relative to the main diagonal entries. Such matrices do arise in practice; large systems of linear equations resulting from numerical discretization of boundary value problems for elliptic partial differential equations can be of this form.

In some differential equations problems involving the long-term stability of an oscillating system, one is sometimes interested in showing that the eigenvalues $\{\lambda_i\}$ of a matrix all lie in the left half-plane, that is, that $\mathrm{Re}(\lambda_i)<0$. And sometimes in statistics or numerical analysis one needs to show that a Hermitian matrix is positive definite, that is, that all $\lambda_i>0$.

Sometimes one wants to locate the eigenvalues of a matrix in a *bounded* set that is easily characterized. We know that all the eigenvalues of a matrix A are located in a disc in the complex plane centered at the origin and having radius $|||A|||$, where $|||\cdot|||$ is any matrix norm. But can one do better than this by more precisely locating regions that must either include or exclude the eigenvalues? We shall see that one can.

Finally, suppose that one knows exactly the eigenvalues of the matrix A, but that A is subjected to a perturbation $A\to A+E$. How do the eigenvalues change? Because the eigenvalues are continuous functions of the entries of A, we have reason to believe that if the perturbation matrix E

is small enough, then the eigenvalues should not change too drastically. But one needs precise bounds to know how small is "small" in each case. The basic issue here is the same as in Section (5.8), where we discussed the sensitivity of the solution of a system of linear equations to perturbations in the data.

6.1 Geršgorin discs

If $A \in M_n$, we can always write $A = D + B$, where $D = \text{diag}(a_{11}, \ldots, a_{nn})$ is just the main diagonal part of A, and B has a zero main diagonal. If we set $A_\epsilon \equiv D + \epsilon B$ for any $\epsilon \in \mathbf{C}$, then $A_0 = D$ and $A_1 = A$. The eigenvalues of $A_0 = D$ are easy to locate: They are just the points $a_{11}, \ldots, a_{nn}$ in the complex plane. We have reason to suspect that if ϵ is small enough, then the eigenvalues of A_ϵ will be located in some small neighborhoods of the points $a_{11}, \ldots, a_{nn}$. The following theorem (often called the Geršgorin disc theorem) makes this observation precise: There are indeed some easily computed discs centered at the points a_{ii} that are guaranteed to contain the eigenvalues.

6.1.1 **Theorem** (Geršgorin). Let $A = [a_{ij}] \in M_n$, and let

$$R_i'(A) \equiv \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \qquad 1 \le i \le n$$

denote the *deleted absolute row sums* of A. Then all the eigenvalues of A are located in the union of n discs

$$\bigcup_{i=1}^{n} \{z \in \mathbf{C}: |z - a_{ii}| \le R_i'(A)\} \equiv G(A) \tag{6.1.2}$$

Furthermore, if a union of k of these n discs forms a connected region that is disjoint from all the remaining $n-k$ discs, then there are precisely k eigenvalues of A in this region.

Proof: Let λ be an eigenvalue of A, and suppose $Ax = \lambda x$, $x = [x_i] \neq 0$ There is an element of x that has largest absolute value, say $|x_p| \ge |x_i|$ for all $i = 1, 2, \ldots, n$, and $x_p \neq 0$. Then the assumption that $Ax = \lambda x$ means that

$$\lambda x_p = [\lambda x]_p = [Ax]_p = \sum_{j=1}^{n} a_{pj} x_j$$

which is equivalent to

$$x_p(\lambda - a_{pp}) = \sum_{\substack{j=1 \\ j \neq p}}^{n} a_{pj} x_j$$

But then the triangle inequality permits us to conclude that

$$|x_p||\lambda - a_{pp}| = \left| \sum_{\substack{j=1 \\ j \neq p}}^{n} a_{pj} x_j \right| \le \sum_{\substack{j=1 \\ j \neq p}}^{n} |a_{pj} x_j| = \sum_{\substack{j=1 \\ j \neq p}}^{n} |a_{pj}||x_j|$$

$$\le |x_p| \sum_{\substack{j=1 \\ j \neq p}}^{n} |a_{pj}| = |x_p| R_p'$$

Thus, $|\lambda - a_{pp}| \le R_p'$ for some p; that is, λ lies in a closed disc around a_{pp} of radius R_p'. Since we do not know *which* p is appropriate to each eigenvalue λ (unless we know the associated eigenvector, in which case we would know λ exactly and would not be interested in locating it), we can only conclude that λ lies in the union of all such discs, which is the region (6.12).

In order to prove the second assertion of the theorem, write $A = D + B$, where $D = \text{diag}(a_{11}, \ldots, a_{nn})$ and set $A_\epsilon \equiv D + \epsilon B$ for $\epsilon \in [0, 1]$. Notice that $R_i'(A_\epsilon) = R_i'(\epsilon B) = \epsilon R_i'(A)$. For convenience, suppose the first k discs

$$\bigcup_{i=1}^{k} \{z \in \mathbf{C}: |z - a_{ii}| \le R_i'\}$$

form a connected region G_k that is disjoint from the complementary region G_k^c consisting of the $n-k$ remaining discs, that is, $G_k^c = G(A) \setminus G_k$. Notice that the union of the first k discs of A_ϵ

$$G_k(\epsilon) \equiv \bigcup_{i=1}^{k} \{z \in \mathbf{C}: |z - a_{ii}| \le R_i'(A_\epsilon) = \epsilon R_i'(A)\}$$

is contained in the connected set $G_k \equiv G_k(1)$ for all $\epsilon \in [0, 1]$, but that $G_k(\epsilon)$ may not itself be a connected set for all such ϵ. Furthermore, none of the complementary regions $G_k^c(\epsilon) \equiv G_n(\epsilon) \setminus G_k(\epsilon)$ ever intersect G_k. For each $i = 1, \ldots, k$, consider the eigenvalues $\lambda_i(A_0) = a_{ii}$ and $\lambda_i(A_\epsilon)$, $\epsilon > 0$. Because the eigenvalues are continuous functions of the entries of A (see Appendix D), and because all $\lambda_i(A_\epsilon) \in G_k(\epsilon) \subset G_k$ for all $\epsilon \in [0, 1]$, each $\lambda_i(A_0)$ is joined to some $\lambda_i(A_1) = \lambda_i(A)$ by the continuous curve in G_k given by $\{\lambda_i(A_\epsilon): 0 \le \epsilon \le 1\}$. For each $\epsilon \in [0, 1]$ we conclude that there are at least k eigenvalues of A_ϵ contained in $G_k(\epsilon)$. But there cannot be more than k because the remaining $n-k$ eigenvalues of A_0 start outside the connected set G_k and follow continuous curves that must remain within the complementary region G_k^c; because of continuity and connectivity (this is the intermediate value theorem for continuous functions), they cannot leap the void between G_k^c and G_k. □

The region $G(A)$ in (6.1.2) is often called the *Geršgorin region* (for rows) of A; the individual discs in $G(A)$ are called *Geršgorin discs,* and

the boundaries of these discs are called *Geršgorin circles.* Since A and A^T have the same eigenvalues, one can obtain a Geršgorin disc theorem for columns by applying the Geršgorin disc theorem to A^T to obtain a region that contains the eigenvalues of A and is specified in terms of deleted absolute column sums

$$C'_j(A) \equiv \sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{ij}|$$

6.1.3 **Corollary.** If $A = [a_{ij}] \in M_n$, then all the eigenvalues of A are located in the union of n discs

$$\bigcup_{j=1}^{n} \{z \in \mathbf{C} \colon |z - a_{jj}| \leq C'_j\} = G(A^T) \tag{6.1.4}$$

Furthermore, if a union of k of these discs forms a connected region that is disjoint from all the remaining $n-k$ discs, then there are precisely k eigenvalues of A in this region.

Exercise. Show that all the eigenvalues of A lie in the intersection of the regions (6.1.2) and (6.1.4), that is, in $G(A) \cap G(A^T)$. Illustrate with the 3-by-3 matrix $[a_{ij}]$ with $a_{ij} = i/j$.

Since all the eigenvalues of A are located in the two regions (6.1.2) and (6.1.4), the largest modulus eigenvalue of A is located there. The point in the ith disc in $G(A)$ that is farthest from the origin has modulus

$$|a_{ii}| + R'_i = \sum_{j=1}^{n} |a_{ij}|$$

so the largest of these values must be an upper bound for the largest modulus eigenvalue of A. Of course, a similar argument can be made for the absolute column sums.

6.1.5 **Corollary.** If $A = [a_{ij}] \in M_n$, then

$$\rho(A) \leq \min\left\{\max_i \sum_{j=1}^{n} |a_{ij}|, \max_j \sum_{i=1}^{n} |a_{ij}|\right\}$$

This result is no surprise, since it says that $\rho(A) \leq \|A\|_\infty$ and $\|A^T\|_\infty$ (the maximum absolute row sum and maximum absolute column sum norms), and this inequality holds for any matrix norm. But it is interesting to have an essentially geometric derivation of this fact.

Since $S^{-1}AS$ has the same eigenvalues as A whenever S is invertible,

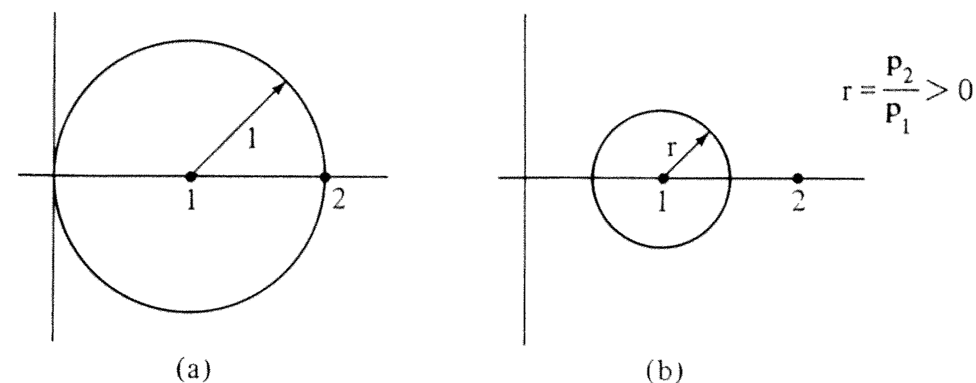


Figure 6.1.7

we can apply the Geršgorin theorem to $S^{-1}AS$; perhaps for some choice of S the bounds obtained may be sharper. A particularly convenient choice is $S = D = \text{diag}(p_1, p_2, \ldots, p_n)$ with all $p_i > 0$. One calculates easily that $D^{-1}AD = [p_j a_{ij}/p_i]$. Applying the Geršgorin theorem to $D^{-1}AD$ and to its transpose yields the following.

6.1.6 **Corollary.** Let $A = [a_{ij}] \in M_n$ and let $p_1, p_2, \ldots, p_n$ be positive real numbers. Then all the eigenvalues of A lie in the region

$$\bigcup_{i=1}^{n} \left\{z \in \mathbf{C} \colon |z - a_{ii}| \leq \frac{1}{p_i} \sum_{\substack{j=1 \\ j \neq i}}^{n} p_j |a_{ij}|\right\} = G(D^{-1}AD)$$

as well as in the region

$$\bigcup_{j=1}^{n} \left\{z \in \mathbf{C} \colon |z - a_{jj}| \leq p_j \sum_{\substack{i=1 \\ i \neq j}}^{n} \frac{1}{p_i} |a_{ij}|\right\} = G[(D^{-1}AD)']$$

The matrix $A = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$ has eigenvalues 1 and 2. A straightforward application of the Geršgorin theorem gives a rather gross estimate for the eigenvalues (Fig. 6.1.7a), but the extra parameters in the last corollary give enough flexibility to obtain an arbitrarily good estimate of the eigenvalues (Fig. 6.1.7b).

Exercise. Consider the matrix

$$A = \begin{bmatrix} 7 & -16 & 8 \\ -16 & 7 & -8 \\ 8 & -8 & -5 \end{bmatrix}$$

Use the Geršgorin theorem to say as much as you can about the location of the eigenvalues of A and the spectral radius of A. Then consider $D^{-1}AD$, where $D = \operatorname{diag}(p_1, p_2, p_3)$, and see if you can obtain any improvement in your location of the eigenvalues. Finally, compute the actual eigenvalues and comment on how well you did with the estimates.

Exercise. Show that every eigenvalue of A lies in the set $\bigcap_D G(D^{-1}AD)$ where the intersection is over all diagonal matrices with positive main diagonal entries.

The idea of introducing free parameters can also be used to obtain a more general form of the estimates (6.1.5) for the spectral radius.

6.1.8 **Corollary.** Let $A = [a_{ij}] \in M_n$. Then

$$\rho(A) \le \min_{p_1,\dots,p_n>0} \max_{1\le i\le n} \frac{1}{p_i} \sum_{j=1}^{n} p_j |a_{ij}|$$

and

$$\rho(A) \le \min_{p_1,\dots,p_n>0} \max_{1\le j\le n} p_j \sum_{i=1}^{n} \frac{1}{p_i} |a_{ij}|$$

Exercise. Prove Corollary (6.1.8).

Exercise. Let $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$, where a, b, c, and d are strictly positive real numbers.

(a) By direct calculation, find an explicit diagonal matrix $\tilde{D}$ such that $\|\tilde{D}^{-1}A\tilde{D}\|_\infty = \min_D \|D^{-1}AD\|_\infty$, where the minimum is taken over all diagonal matrices D with positive main diagonal entries.
(b) Calculate $\|\tilde{D}^{-1}A\tilde{D}\|_\infty \equiv r$.
(c) Calculate $\rho(A)$ explicitly.
(d) Note that $r = \rho(A)$.

We shall show later that if A is any n-by-n positive matrix (or, more generally, is irreducible and nonnegative), then the minimum over all D of the maximum row sum of $D^{-1}AD$ is always equal to the spectral radius. This is not the case in general.

Exercise. Consider $A = \left[\begin{smallmatrix} 1 & 1 \\ -1.5 & 2 \end{smallmatrix}\right]$ and show that $\rho(A) < \min \|D^{-1}AD\|_\infty$ over all $D = \operatorname{diag}(p_1, p_2)$ with p_1 and p_2 positive.

If one has some additional information about a matrix, which forces its eigenvalues to lie in (or not in) certain sets, then this information can be used along with the Geršgorin theorem to give an even more precise location for the eigenvalues. For example, if A is Hermitian, then the eigenvalues of A must all be real so they must lie in the set $\mathbf{R} \cap G(A)$, which is a finite union of closed real intervals.

Exercise. What can be said about the location of the eigenvalues of a skew-Hermitian matrix? A unitary matrix? A real orthogonal matrix?

Since a matrix is invertible if and only if 0 is not an eigenvalue, it is of interest to develop conditions that exclude the origin from the region known to contain the eigenvalues.

6.1.9 **Definition.** Let $A = [a_{ij}] \in M_n$. The matrix A is said to be *diagonally dominant* if

$$|a_{ii}| \ge \sum_{\substack{j=1 \\ j\ne i}}^{n} |a_{ij}| = R_i' \quad \text{for all} \quad i = 1, \dots, n$$

It is said to be *strictly diagonally dominant* if

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\ne i}}^{n} |a_{ij}| = R_i' \quad \text{for all} \quad i = 1, \dots, n$$

From the geometry of the situation it is apparent that 0 cannot lie in any closed Geršgorin disc if A is strictly diagonally dominant. Furthermore, if all the main diagonal entries a_{ii} are real and positive, then each disc actually lies in the open right half-plane; if A is Hermitian as well, then the eigenvalues must all be positive. We summarize these observations in the following theorem, of which part (a) is known as the Levy–Desplanques theorem [see Corollary (5.6.17)].

6.1.10 **Theorem.** Let $A = [a_{ij}] \in M_n$ be strictly diagonally dominant. Then

(a) A is invertible.
(b) If all main diagonal entries of A are positive, then all the eigenvalues of A have positive real part.
(c) If A is Hermitian and all main diagonal entries of A are positive, then all the eigenvalues of A are real and positive.

Exercise. Consider $\left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 1 & 1 \\ 1-\epsilon & 1 \end{smallmatrix}\right]$ to show that diagonal dominance alone is not sufficient to guarantee invertibility, and that strict diagonal dominance is not necessary for invertibility.

By using the extra parameters in Corollary (6.1.6), the assumption of strict diagonal dominance as a sufficient condition for invertibility can be relaxed slightly.

6.1.11 Theorem. Let $A=[a_{ij}]\in M_n$ have all diagonal entries nonzero and be diagonally dominant with $|a_{ii}|>R_i'$ for all but one value of $i=1,\dots,n$. Then A is invertible.

Proof: The hypothesis is that for some k, $|a_{kk}|=R_k'$ and $|a_{ii}|>R_i'$ for all $i\neq k$. In (6.1.6), let $p_i=1$ for all $i\neq k$ and let $p_k=1+\epsilon$, $\epsilon>0$. Then

$$\frac{1}{p_k}\sum_{\substack{j=1\\ j\neq k}}^{n} p_j|a_{ij}|=\frac{1}{1+\epsilon}R_k'<|a_{kk}| \quad \text{for any} \quad \epsilon>0$$

and

$$\frac{1}{p_i}\sum_{\substack{j=1\\ j\neq i}}^{n} p_j|a_{ij}|=R_i'+\epsilon|a_{ik}| \quad \text{for all} \quad i\neq k$$

But since $R_i'<|a_{ii}|$ for all $i\neq k$, we can choose $\epsilon>0$ small enough so that $R_i'+\epsilon|a_{ik}|<|a_{ii}|$ for all $i\neq k$. By Corollary (6.1.6), the point $z=0$ is therefore excluded from $G(D^{-1}AD)$ and so A must be invertible. □

The Geršgorin theorem and its variations give inclusion regions for the eigenvalues of A, which depend only on the main diagonal entries of A and the *absolute values* of the off-diagonal entries. Using the fact that $S^{-1}AS$ has the same eigenvalues as A led us to (6.1.6) and to the fact that the closed set

$$\bigcap_D G(D^{-1}AD), \qquad D=\operatorname{diag}(p_1,\dots,p_n), \qquad \text{all } p_i>0 \tag{6.1.12}$$

contains all the eigenvalues of $A\in M_n$. We know that we could get even smaller inclusion regions for the eigenvalues if we were to admit more complicated similarities than diagonal ones, but if we restrict ourselves to just diagonal similarities and to the use of just the main diagonal entries and the absolute values of the off-diagonal entries, can we somehow do better than (6.1.12)?

The answer is no, for the following reason: Let z be any given point on the boundary of the set (6.1.12). Then R. Varga has shown that there exists a matrix $B=[b_{ij}]\in M_n$ such that $b_{ii}=a_{ii}$ for all $i=1,\dots,n$ and $|b_{ij}|=|a_{ij}|$ for all $i,j=1,\dots,n$ and such that z is an eigenvalue of B.

Problems

1. Consider the following iterative algorithm for solving the n-by-n system of linear equations $Ax=y$, where A and y are given:

(i) Define $B\equiv I-A$ and rewrite the system as $x=Bx+y$.
(ii) Choose an initial approximation $x^{(0)}$ to the solution in any way you wish.
(iii) For $m=0,1,2,\dots$ calculate $x^{(m+1)}=Bx^{(m)}+y$.
(iv) Hope that $x^{(m)}\to x$ (the solution) as $m\to\infty$.

(a) Denote by $\epsilon^{(m)}=x^{(m)}-x$ the error in the mth approximation to the solution, and show that $\epsilon^{(m)}=B^m(x^{(0)}-x)$. (b) Conclude that if $\rho(I-A)<1$, then this algorithm works in the sense that $x^{(m)}\to x$ as $m\to\infty$ regardless of the choice of the initial approximation $x^{(0)}$. (c) Use the Geršgorin theorem to give a simple explicit condition on A that is sufficient for this algorithm to work.

2. Show that $\bigcap_S G(S^{-1}AS)=\sigma(A)$ if the intersection is taken over all nonsingular S.

3. Use (6.1.5) to show that

$$|\det A|\leq\prod_{i=1}^{n}\left(\sum_{j=1}^{n}|a_{ij}|\right)$$

for any $A\in M_n$, with a similar inequality for the columns. *Hint:* If any row of A is zero there is nothing to prove. If all rows of A are nonzero, then let B be the matrix whose rows consist of the rows of A divided by the corresponding absolute row sums of A. Then $\rho(B)\leq 1$ by (6.1.5), so $|\det B|\leq 1$. Notice that this says that

$$|\det A|\leq\prod_{i=1}^{n}\|a_i\|_1$$

where the vectors a_i are the respective rows (or columns) of A. Is there such an inequality for other norms? *Hint:* See (7.8.2).

4. In the text we derived Theorem (6.1.10a) – the Levy-Desplanques theorem – from the Geršgorin theorem (6.1.1). Show that the first part of (6.1.1) [the fact that the region (6.1.2) contains all the eigenvalues of A] follows from part (a) of (6.1.10). *Hint:* Apply (6.1.10a) to the matrix $\lambda I-A$.

5. Suppose that $A\in M_n$ is a real matrix whose n Geršgorin discs are all mutually disjoint. Show that all the eigenvalues of A are real. More generally, show that the same is true, and for the same reason, if a complex matrix $A\in M_n$ has real main diagonal entries and its characteristic polynomial has only real coefficients.

6. Show that if $A=[a_{ij}]\in M_n$ and if $|a_{ii}|>R_i'$ for k different values of i, then $k\leq\operatorname{rank} A$.

7. Suppose that $A\in M_n$ is idempotent ($A^2=A$), but that $A\neq I$. Show

that A cannot be strictly diagonally dominant [or irreducibly diagonally dominant; see (6.2.25) and (6.2.27)].

8. Suppose that $A \in M_n$ is strictly diagonally dominant, $|a_{ii}| > R_i'$ for all $i = 1, \ldots, n$. Show that $|a_{kk}| > C_k'$ for at least one value of $k = 1, \ldots, n$.

9. Suppose $A = [a_{ij}] \in M_n$ is strictly diagonally dominant, and let $D = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn})$. Show that D is invertible and that $\rho(I - D^{-1}A) < 1$. *Hint:* Use Corollary (6.1.5).

10. If $A = [a_{ij}] \in M_n$ and if $R_i = R_i' + |a_{ii}|$ denotes the sum of the absolute values of all the entries in the ith row of A, show that

$$\operatorname{rank} A \geq \sum_{i=1}^{n} \frac{|a_{ii}|}{R_i}$$

where we agree that $0/0 = 0$ in this sum. *Hint:* Multiplying all the elements in a row by the same nonzero scalar does not change the rank, so it suffices to assume that all $a_{ii} \geq 0$ and all R_i are either zero or 1. In this case, all the eigenvalues of A lie in the unit disc and one must show that

$$\operatorname{rank} A \geq \sum_{i=1}^{n} a_{ii}$$

Show that $\sum a_{ii} = \operatorname{tr} A = \sum \lambda_i \leq \sum |\lambda_i| \leq$ number of nonzero eigenvalues of $A \leq \operatorname{rank} A$.

11. If $A = [a_{ij}] = [a_1\, a_2 \ldots a_n] \in M_n$, show that

$$\operatorname{rank} A \geq \sum_{i=1}^{n} \frac{|a_{ii}|^2}{\|a_i\|_2^2}$$

where we agree that $0/0 = 0$ in the sum. *Hint:* As in Problem 10, show that it suffices to consider the case in which all the columns of A have unit Euclidean length, that is, all $\|a_i\|_2 = 1$, and in this case one must show that

$$\operatorname{rank} A \geq \sum_{i=1}^{n} |a_{ii}|^2 = \sum_{i=1}^{n} |e_i^* a_i|^2$$

where $\{e_1, e_2, \ldots, e_n\}$ is the standard orthonormal basis of $\mathbf{C}^n$. If A has rank k, show that there are k orthonormal vectors $v_1, \ldots, v_k \in \mathbf{C}^n$ such that $\operatorname{Span}\{v_1, \ldots, v_k\} = \operatorname{Span}\{a_1, \ldots, a_n\}$. Then

$$a_i = \sum_{j=1}^{k} (v_j^* a_i) v_j, \qquad \text{so} \quad e_i^* a_i = \sum_{j=1}^{k} (v_j^* a_i)(e_i^* v_j)$$

and

$$\begin{aligned}\sum_{i=1}^{n} |e_i^* a_i|^2 &\leq \sum_{i=1}^{n} \left[\left(\sum_{j=1}^{k} |v_j^* a_i|^2\right)\left(\sum_{j=1}^{k} |e_i^* v_j|^2\right)\right] \\ &= \sum_{j=1}^{k} \sum_{i=1}^{n} |e_i^* v_j|^2 = \sum_{j=1}^{k} 1 \\ &= k = \operatorname{rank} A\end{aligned}$$

Further Readings. A discussion of the Geršgorin theorem with some numerical examples can be found in [Ste]. The original reference is S. Geršgorin, "Über die Abgrenzung der Eigenwerte einer Matrix," *Izv. Akad. Nauk. S.S.S.R.* 7 (1931), 749–754. There is a generalization of Geršgorin's theorem (6.1.1) that gives inclusion regions for the spectrum of the generalized eigenvalue problem $Ax = \lambda Bx$ and covers the case in which B is singular; see G. W. Stewart, "Gerschgorin Theory for the Generalized Eigenvalue Problem," *Math. Comput.* 29 (1975), 600–606. For a proof that the region (6.1.12) has the optimality property stated in the last paragraph of the section, see R. Varga, "Minimal Gerschgorin Sets," *Pacific J. Math.* 15 (1965), 719–729.

6.2 Geršgorin discs – a closer look

We have seen that strict diagonal dominance is sufficient for invertibility but that diagonal dominance is not. Consideration of some 2-by-2 examples suggests the conjecture that diagonal dominance together with strict inequality

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \quad \text{for at least one value of } i = 1, \ldots, n \tag{6.2.1}$$

may be sufficient for invertibility. Unfortunately, this is not the case, as is shown by the example

$$\begin{bmatrix} 4 & 2 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \tag{6.2.2}$$

However, there are useful conditions on a diagonally dominant matrix under which (6.2.1) *is* sufficient to guarantee invertibility, and they lead to some very interesting ideas in graph theory. The fundamental observation is that if A is diagonally dominant, then 0 cannot be an interior point of any individual Geršgorin disc.

Exercise. Show that a given point λ is not an interior point of any Geršgorin disc of A if and only if

$$|\lambda - a_{ii}| \geq R_i' = \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \quad \text{for all } i=1,\ldots,n \tag{6.2.2a}$$

Show that every point λ on the boundary of $G(A)$ satisfies these inequalities. Consider $\lambda = 0$ and $A = \begin{bmatrix}1&1\\1&i\end{bmatrix} \oplus \begin{bmatrix}-1&1\\1&-i\end{bmatrix}$ to show that an interior point of $G(A)$ can also satisfy the inequalities (6.2.2a).

A careful analysis of the proof of Theorem (6.1.1) clarifies what happens when an eigenvalue of A satisfies the inequalities (6.2.2a).

6.2.3 **Lemma.** Let $A = [a_{ij}] \in M_n$ and let λ be an eigenvalue of A that satisfies the inequalities (6.2.2a). Let $Ax = \lambda x$, $x = [x_i] \neq 0$, and suppose p is an index such that $|x_p| = \max_{1\leq i\leq n}|x_i| = \|x\|_\infty \neq 0$. Then

(a) If k is any index such that $|x_k| = |x_p|$, then $|\lambda - a_{kk}| = R_k'$; that is, the kth Geršgorin circle passes through λ; and
(b) If $|x_k| = |x_p|$ for some $k = 1,\ldots,n$ and if $a_{kj} \neq 0$ for some $j \neq k$, then $|x_j| = |x_p|$ as well.

Proof: Just as in the proof of the Geršgorin theorem, we have

$$(\lambda - a_{ii})x_i = \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}x_j \quad \text{for all } i=1,\ldots,n$$

and hence

$$\begin{aligned}|\lambda - a_{ii}||x_i| &= \left|\sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}x_j\right| \leq \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}x_j| = \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}||x_j| \\ &\leq \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}||x_p| = R_i'|x_p|\end{aligned} \tag{6.2.4}$$

Thus, if k is any index such that $|x_k| = |x_p|$, we must have $|\lambda - a_{kk}| \leq R_k'$. But the hypothesis is that $|\lambda - a_{ii}| \geq R_i'$ for all $i = 1,\ldots,n$, so we must actually have *equality* in both of the inequalities in (6.2.4) for $i = k$; that is,

$$|\lambda - a_{kk}||x_k| = \sum_{\substack{j=1\\ j\neq k}}^{n} |a_{kj}||x_j| = \sum_{\substack{j=1\\ j\neq k}}^{n} |a_{kj}||x_k| = R_k'|x_k| \tag{$\dagger$}$$

Since $|x_k| = \|x\|_\infty > 0$, assertion (a) follows from the identity

$$|\lambda - a_{kk}||x_k| = R_k'|x_k|$$

Assertion (b) follows from the center identity in ($\dagger$)

$$\sum_{\substack{j=1\\ j\neq k}}^{n} |a_{kj}|(|x_k| - |x_j|) = 0$$

because every term in this sum must be nonnegative. □

This lemma looks rather technical, but it has as an immediate consequence the following useful result and its corollary.

6.2.5 **Theorem.** Let $A \in M_n$, and let λ be an eigenvalue of A that is a boundary point of $G(A)$ or, more generally, satisfies the inequalities (6.2.2a). Suppose that all the entries of A are nonzero. Then

(a) Every Geršgorin circle of A passes through λ; and
(b) If $Ax = \lambda x$, $x = [x_i] \neq 0$, then $|x_i| = |x_j|$ for all $i, j = 1,\ldots,n$.

Exercise. Deduce Theorem (6.2.5) from Lemma (6.2.3).

6.2.6 **Corollary.** Let $A = [a_{ij}] \in M_n$, and suppose that all the entries of A are nonzero. If A is diagonally dominant and if $|a_{ii}| > R_i'$ for at least one value of $i = 1,\ldots,n$, then A is invertible.

Proof: If A were not invertible, then 0 would be an eigenvalue of A. Since A is diagonally dominant, 0 cannot be an interior point of any Geršgorin disc and hence $\lambda = 0$ satisfies the inequalities (6.2.2a). The theorem says that every Geršgorin circle must pass through 0, but if $|a_{ii}| > R_i'$, then the ith circle cannot pass through 0. □

The previous result is both useful and interesting, but we can do much better (with regard to the assumption on zero entries in A) if we use more carefully the information in Lemma (6.2.3).

6.2.7 **Definition.** A matrix $A = [a_{ij}] \in M_n$ is said to have *property SC* if for every pair of distinct integers p, q with $1 \leq p$, $q \leq n$ there is a sequence of distinct integers $k_1 = p$, $k_2, k_3, \ldots, k_{m-1}, k_m = q$, $1 \leq m \leq n$, such that all of the matrix entries $a_{k_1k_2}, a_{k_2k_3}a_{k_{m-1}k_m}$ are nonzero.

For example, the matrix (6.2.2) does not have property *SC* because the pair 2, 1 does not admit such a sequence of nonzero entries. The pair 1, 2 does admit such a sequence, however.

Using this notion and Lemma (6.2.3), we can obtain the following improvement on (6.2.5).

6.2.8 Better theorem. Let $A=[a_{ij}]\in M_n$, and suppose that λ is an eigenvalue of A that is a boundary point of $G(A)$, or, more generally, satisfies the inequalities (6.2.2a). If A has property *SC*, then

(a) Every Geršgorin circle passes through λ; and
(b) If $Ax=\lambda x$ and $x=[x_i]\neq 0$, then $|x_i|=|x_j|$ for all $i,j=1,\ldots,n$.

Proof: Let $Ax=\lambda x$ with $|x_i|\leq|x_p|=\|x\|_\infty>0$ for all $i=1,\ldots,n$. Then $|\lambda-a_{pp}|=R_p'$ by Lemma (6.2.3). Let q be any other index, $1\leq q\leq n$, $q\neq p$. Since A has property *SC*, there is a sequence of distinct indices $k_1=p,k_2,k_3,\ldots,k_m=q$ such that all of the matrix entries $a_{k_1k_2},\ldots,a_{k_{m-1}k_m}$ are nonzero. Since $a_{k_1k_2}=a_{pk_2}\neq 0$, we see by assertion (b) of (6.2.3) that $|x_p|=|x_{k_2}|$. But then $a_{k_2k_3}\neq 0$ and so $|x_{k_3}|=|x_{k_2}|=|x_p|$. Proceeding in this way we conclude that $|x_{k_i}|=|x_p|$ for all $i=1,\ldots,m$ and hence [by (6.2.3)(a)], $|\lambda-a_{k_mk_m}|=|\lambda-a_{qq}|=R_q'$; that is, the qth Geršgorin circle passes through λ and $|x_q|=|x_p|$. But since q was an arbitrary index, we conclude that every Geršgorin circle passes through λ and that all $|x_i|=|x_p|$, $i=1,\ldots,n$. □

Just as in (6.2.6), we can deduce a useful sufficient condition for invertibility from this result.

6.2.9 Better corollary. Let $A=[a_{ij}]\in M_n$ and suppose that A has property *SC*. If A is diagonally dominant and if $|a_{ii}|>R_i'$ for at least one value of $i=1,\ldots,n$, then A is invertible.

Exercise. Deduce (6.2.9) from (6.2.8).

Exercise. Show that the matrix (6.2.2) does not have property *SC*.

What is this strange property *SC*? Notice that it has to do only with the locations of the off-diagonal nonzero entries of A – the main diagonal entries and the precise values of the off-diagonal entries are irrelevant. Because of this observation, we define two matrices related to A.

6.2.10 Definition. If $A=[a_{ij}]\in M_{m,n}$ we set $|A|\equiv[|a_{ij}|]$ and $M(A)\equiv[\mu_{ij}]$, where $\mu_{ij}=1$ if $a_{ij}\neq 0$ and $\mu_{ij}=0$ if $a_{ij}=0$. The matrix $M(A)$ is called the *indicator matrix* of A.

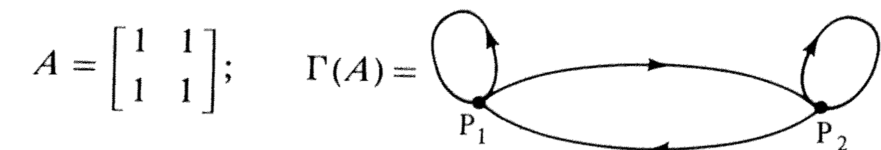
Exercise. Show that a matrix $A\in M_n$ has property *SC* if and only if either (and hence both) $|A|$ or $M(A)$ has property *SC*.

The concept of a sequence of nonzero entries of A that arises in the statement of property *SC* can be summarized visually in terms of certain paths in a graph associated with A.

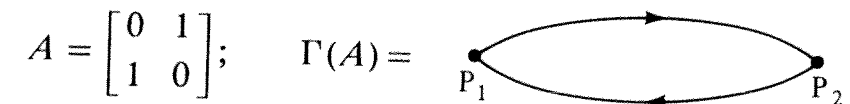
6.2.11 Definition. The *directed graph of* $A\in M_n$, denoted by $\Gamma(A)$, is the directed graph on n nodes $P_1,P_2,\ldots,P_n$ such that there is a directed arc in $\Gamma(A)$ from P_i to P_j if and only if $a_{ij}\neq 0$ $(\mu_{ij}\neq 0)$.

Examples

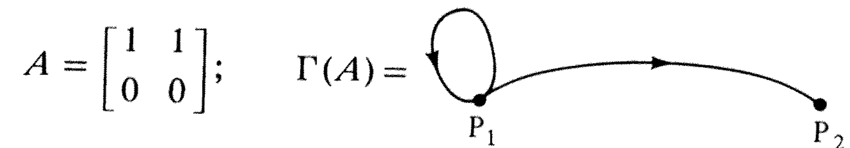
$A=\begin{bmatrix}1&1\\1&1\end{bmatrix}$; $\Gamma(A)=$



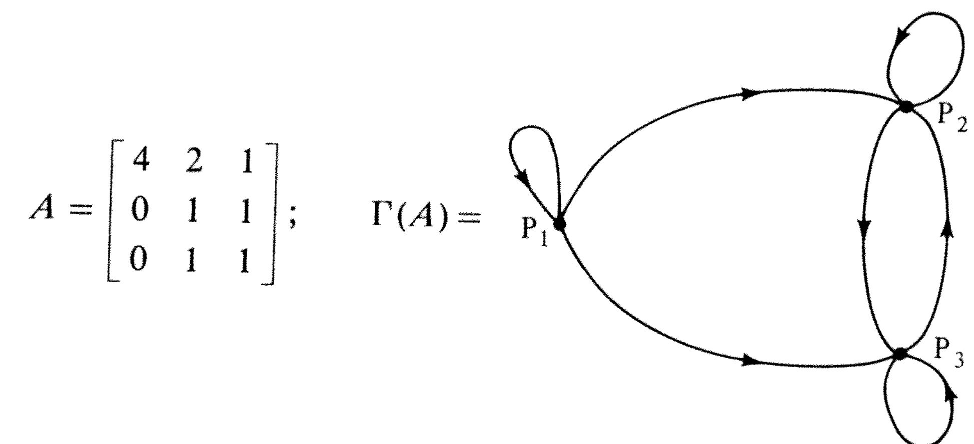
$A=\begin{bmatrix}0&1\\1&0\end{bmatrix}$; $\Gamma(A)=$



$A=\begin{bmatrix}1&1\\0&0\end{bmatrix}$; $\Gamma(A)=$



$A=\begin{bmatrix}4&2&1\\0&1&1\\0&1&1\end{bmatrix}$; $\Gamma(A)=$



6.2.12 Definition. A *directed path* γ in a graph Γ is a sequence of arcs $P_{i_1}P_{i_2},P_{i_2}P_{i_3},P_{i_3}P_{i_4},\ldots$ in Γ. The *ordered list of nodes* in the directed path γ is $P_{i_1},P_{i_2},\ldots$. The *length* of a directed path is the number of successive arcs in the directed path if this number is finite; otherwise, the directed path is said to have infinite length. A *cycle* is a directed path that begins and ends at the same node; this node occurs exactly twice in the

ordered list of nodes in the path, and no other node occurs more than once in the list; some authors would call this a *simple directed cycle.* A cycle of length 1 is called a *loop* or *trivial cycle.*

6.2.13 **Definition.** A directed graph Γ is *strongly connected* if between every pair of distinct nodes P_i, P_j in Γ there is a directed path of finite length that begins at P_i and ends at P_j.

6.2.14 **Theorem.** Let $A \in M_n$. Then A has property SC if and only if the directed graph $\Gamma(A)$ is strongly connected.

Exercise. Prove the theorem.

Exercise. Show that Γ is strongly connected if it has the property that every pair of nodes belongs to at least one cycle, but that the converse is not correct. *Hint:*

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

There may be more than one directed path between two nodes of a directed graph, but two such paths with different lengths may not be essentially different; one may contain repetitions of one or more subpaths. It is clear that if one ever visits a given node twice in going along a directed path, then the directed path may be shortened (and the end points will be unaffected) by deleting all the intermediate arcs between the first and second visits to the node (the subgraph deleted is, or contains, a cycle).

6.2.15 **Observation.** Let Γ be a directed graph on n nodes. If there is a directed path in Γ between two given nodes, then between these nodes there is a directed path that has length not greater than $n-1$.

How can one tell if a given matrix A has property SC? This is equivalent to checking whether $\Gamma(A)$ is strongly connected. If n is not large or if $M(A)$ has a special structure, then one can just inspect $\Gamma(A)$ and trace out paths between all possible pairs of nodes. However, this is not practical in general, so we need some explicit computational method.

6.2.16 **Theorem.** Let $A \in M_n$ be given, and let P_i and P_j be given nodes of $\Gamma(A)$. There exists a directed path of length m in $\Gamma(A)$ from P_i to P_j if and only if $(|A|^m)_{ij} \neq 0$ or, equivalently, $[M(A)^m]_{ij} \neq 0$.

Proof: We proceed by induction. For $m=1$ the assertion is trivial. For $m=2$ we compute

$$[|A|^2]_{ij} = \sum_{k=1}^{n} [|A|]_{ik}[|A|]_{kj} = \sum_{k=1}^{n} |a_{ik}|\,|a_{kj}|$$

so that $[|A|^2]_{ij} \neq 0$ if and only if for at least one value of k, a_{ik} and a_{kj} are *both* nonzero. But this is the case if and only if there exists a path of length 2 in $\Gamma(A)$ from P_i to P_j. In general, suppose the assertion has been proved for $m=q$. Then

$$[|A|^{q+1}]_{ij} = \sum_{k=1}^{n} [|A|^q]_{ik}[|A|]_{kj} = \sum_{k=1}^{n} [|A|^q]_{ik}\,|a_{kj}| \neq 0$$

if and only if for at least one value of k, $[|A|^q]_{ik}$ and $|a_{kj}|$ are both nonzero. This is equivalent to having a path from P_i to P_k of length q and one from P_k to P_j of length 1, and this is the case if and only if there is a path from P_i to P_j of length $q+1$. The same argument works for $M(A)$. □

6.2.17 **Definition.** Let $A=[a_{ij}] \in M_n$. We say that $A \geq 0$ (A is *nonnegative*) if all its entries a_{ij} are real and nonnegative. We say that $A > 0$ (A is *positive*) if all its entries a_{ij} are real and positive.

6.2.18 **Corollary.** Let $A \in M_n$. Then $|A|^m > 0$ if and only if from each node P_i to each node P_j in $\Gamma(A)$ there is a directed path in $\Gamma(A)$ of length exactly m. The same is true for $M(A)^m$.

6.2.19 **Corollary.** Let $A \in M_n$. Then A has property SC if and only if $(I+|A|)^{n-1} > 0$ or, equivalently, if $[I+M(A)]^{n-1} > 0$.

Proof: $(I+|A|)^{n-1} = I + (n-1)|A| + \binom{n-1}{2}|A|^2 + \cdots + \binom{n-1}{n-2}|A|^{n-1} > 0$ if and only if for each pair (i,j) of nodes with $i \neq j$ at least one of the terms $|A|, |A|^2, \ldots, |A|^{n-1}$ has a positive (i,j) entry. But Theorem (6.2.16) says this happens if and only if there is some directed path in $\Gamma(A)$ from P_i to P_j. This is equivalent to $\Gamma(A)$ being strongly connected, which is equivalent to A having property SC. □

Exercise. Prove the assertion in Corollary (6.2.19) involving $M(A)$.

6.2.20 **Corollary.** There is a path in $\Gamma(A)$ from P_i to P_j, $i \neq j$, if and only if $[(I+|a|)^{n-1}]_{ij} \neq 0$.

Exercise. Use Corollary (6.2.19) to give an explicit computational test for property *SC* that involves only about $\log_2(n-1)$ matrix multiplications instead of $n-2$ matrix multiplications. *Hint:* Consider $(I+|A|)^2$, the square of this, and so on.

Before leaving this subject, we introduce one more equivalent characterization of property *SC*. It is based on the fact that strong connectivity of $\Gamma(A)$ is just a topological property of $\Gamma(A)$ – it has nothing to do with the labeling assigned to the nodes of $\Gamma(A)$. If we permute the labels of the nodes, the graph stays either strongly connected or not strongly connected. Notice that if we interchange the ith and jth rows of A as well as the ith and jth columns, this has the effect on $\Gamma(A)$ of interchanging the labels on nodes P_i and P_j, and vice versa.

Recall that a *permutation matrix* P is a square matrix, all of whose entries are 0 or 1; in each row and column of P there is precisely one 1. Clearly, such a matrix is unitary, hence orthogonal, so $P^T=P^{-1}$. The simplest permutation matrix P has $p_{ij}=p_{ji}=1$ for some fixed choice of i, j and has all other nondiagonal entries 0. The similarity P^TAP then has the effect of interchanging the ith and jth columns of A as well as interchanging the ith and jth rows of A. Any permutation of the rows and columns of A can be obtained as a succession of such interchanges, and any permutation matrix is a finite product of such simple permutation matrices. Thus, if P is a permutation matrix, the similarity P^TAP is obtained from A by a suitable permutation of the rows and columns of A. It is important to know whether some permutation of the rows and columns of A can be found that brings A into the following special block form.

6.2.21 **Definition.** A matrix $A\in M_n$ is said to be *reducible* if either

(a) $n=1$ and $A=0$; or

(b) $n\geq 2$, there is a permutation matrix $P\in M_n$, and there is some integer r with $1\leq r\leq n-1$ such that

$$P^TAP=\begin{bmatrix} B & C \\ 0 & D \end{bmatrix}$$

where $B\in M_r$, $D\in M_{n-r}$, $C\in M_{r,n-r}$, and $0\in M_{n-r,r}$ is a zero matrix.

Note that we do not insist that the blocks B, C, and D have nonzero entries, but only that we should be able to get an $(n-r)$-by-r block of 0 entries in the indicated position by some sequence of row and column interchanges. If $|A|>0$, clearly A is not reducible, and if A is reducible, it must have at least $(n-1)$ 0 entries.

Remark: Suppose we want to solve the system of linear equations $Ax=y$, and suppose that A is reducible. Then if we write $\tilde{A}=P^TAP=\left[\begin{smallmatrix} B & C \\ 0 & D\end{smallmatrix}\right]$, we have $Ax=P\tilde{A}P^Tx=y$, or $\tilde{A}(P^Tx)=P^Ty$. Set $P^Tx=\tilde{x}=[z^T\vdots\zeta^T]^T$ (unknown) and $P^Ty=\tilde{y}=[w^T\vdots\omega^T]^T$ (known), where $z, w\in\mathbf{C}^r$ and $\zeta,\omega\in\mathbf{C}^{n-r}$. Then the system of equations to be solved is equivalent to $\tilde{A}\tilde{x}=\tilde{y}=\left[\begin{smallmatrix} B & C \\ 0 & D\end{smallmatrix}\right]\left[\begin{smallmatrix} z \\ \zeta\end{smallmatrix}\right]=\left[\begin{smallmatrix} w \\ \omega\end{smallmatrix}\right]$, that is, to

$$Bz+C\zeta=w$$
$$D\zeta=\omega$$

If we solve $D\zeta=\omega$ first for ζ, then use ζ in the first equation and solve $Bz=w-C\zeta$ for z, we have *reduced* the original problem to two smaller problems that should, in principle, be easier to solve. It is this observation that motivates the term *reducible.*

6.2.22 **Definition.** A matrix $A\in M_n$ is said to be *irreducible* if it is not reducible.

6.2.23 **Theorem.** A matrix $A\in M_n$ is irreducible if and only if

$$(I+|A|)^{n-1}>0$$

or, equivalently, if $[I+M(A)]^{n-1}>0$.

Proof: We shall actually prove that A is reducible if and only if $(I+|A|)^{n-1}$ has at least one 0 entry. Suppose first that A is reducible and that for some permutation matrix P we have

$$A=P\begin{bmatrix} B & C \\ 0 & D \end{bmatrix}P^T=P\tilde{A}P^T$$

where B, C, 0, and D are block matrices as in Definition (6.2.21). Notice that $|A|=|P\tilde{A}P^T|=P|\tilde{A}|P^T$ since the effect of P is only to permute rows and columns; also notice that $|\tilde{A}|^2, |\tilde{A}|^3, \ldots, |\tilde{A}|^{n-1}$ all have the same $(n-r)$-by-r block of 0's in the lower left corner as $\tilde{A}$. Thus

$$(I+|A|)^{n-1}=(I+P|\tilde{A}|P^T)^{n-1}=(P[I+|\tilde{A}|]P^T)^{n-1}=P(I+|\tilde{A}|)^{n-1}P^T$$
$$=P\left[I+(n-1)|\tilde{A}|+\binom{n-1}{2}|\tilde{A}|^2+\cdots+\binom{n-1}{n-1}|\tilde{A}|^{n-1}\right]P^T$$

and all of the terms in the square brackets have an $(n-r)$-by-r block of 0's in the lower left corner. Thus, $(I+|A|)^{n-1}$ is reducible and hence it cannot have all nonzero entries.

Conversely, suppose for some $p \neq q$ that the (p,q) entry of $(I+|A|)^{n-1}$ is 0. Then we know that there is no directed path in $\Gamma(A)$ from P_p to P_q. Define the set of nodes

$$S_1 \equiv \{P_i : P_i = P_q \text{ or there is a path in } \Gamma(A) \text{ from } P_i \text{ to } P_q\}$$

and let S_2 contain all nodes of $\Gamma(A)$ that are not in S_1. Notice that $S_1 \cup S_2 = \{P_1, \ldots, P_n\}$ and $P_q \in S_1 \neq \varnothing$, so $S_2 \neq \{P_1, \ldots, P_n\}$. If there were a path from some node P_i of S_2 to some node P_j of S_1, then (by definition of S_1) there would be a path from P_i to P_q and so P_i would already be in S_1. Thus, there can be no paths *from* any node of S_2 *to* any node of S_1. Now relabel the nodes so that $S_1 = \{\tilde{P}_1, \ldots, \tilde{P}_r\}$ and $S_2 = \{\tilde{P}_{r+1}, \ldots, \tilde{P}_n\}$ and notice that

$$\tilde{A} = P^T A P = \begin{bmatrix} B & C \\ 0 & D \end{bmatrix}, \qquad B \in M_r, \qquad 0 \in M_{n-r,r}$$

so that A is reducible. The argument for $[I+M(A)]^{n-1} > 0$ is just the same. □

Let us summarize:

6.2.24 Theorem. Let $A \in M_n$. The following are equivalent:

(a) A is irreducible;
(b) $(I+|A|)^{n-1} > 0$;
(c) $[I+M(A)]^{n-1} > 0$;
(d) $\Gamma(A)$ is strongly connected; and
(e) A has property *SC*.

6.2.25 Definition. Let $A \in M_n$. We say that A is *irreducibly diagonally dominant* if

(a) A is irreducible;
(b) A is diagonally dominant, that is, $|a_{ii}| \geq R_i'(A)$ for all $i = 1, \ldots, n$; and
(c) For at least one value of i we have $|a_{ii}| > R_i'(A)$.

Exercise. Show by example that a matrix can be irreducible and diagonally dominant without being irreducibly diagonally dominant.

In our present language, we can rephrase our "better theorem" (6.2.8) and its corollary as follows:

6.2.26 Theorem. Let $A \in M_n$ be irreducible. A boundary point λ of the Geršgorin region $G(A)$ [or, more generally, a point λ satisfying the inequalities (6.2.2a)] can be an eigenvalue of A only if every Geršgorin circle passes through λ.

6.2.27 Corollary (Taussky). Let $A = [a_{ij}] \in M_n$ be irreducibly diagonally dominant. Then

(a) A is invertible;
(b) If all $a_{ii} > 0$, then $\text{Re}(\lambda_i) > 0$ for all eigenvalues λ_i of A; and
(c) If A is Hermitian (or, more generally, if A has only real eigenvalues), and if all main diagonal entries of A are strictly positive, then all the eigenvalues of A are strictly positive.

6.2.28 Corollary. Let $A \in M_n$ be irreducible, and suppose for at least one value of i that

$$R_i = \sum_{j=1}^{n} |a_{ij}| < \|A\|_\infty$$

that is, not all absolute row sums equal the maximum absolute row sum. Then $\rho(A) < \|A\|_\infty$. More generally, if $p_1, \ldots, p_n > 0$, if

$$D = \text{diag}(p_1, p_2, \ldots, p_n)$$

and if $R_i(D^{-1}AD) < \|D^{-1}AD\|_\infty$ for at least one value of i, then $\rho(A) < \|D^{-1}AD\|_\infty$.

Proof: We always have the bound $\rho(A) \leq \|A\|_\infty$ and we have equality if and only if there is some eigenvalue λ of A with $|\lambda| = \|A\|_\infty$. But then by Theorem (6.2.26) every Geršgorin circle must pass through λ. Our assumption that some $R_i < \|A\|_\infty$ prevents this, however. The second assertion follows from the same argument applied to $D^{-1}AD$. □

Problems

1. Show that an irreducible matrix cannot have a 0 row or column.

2. Show by an example that the hypothesis of irreducibility in Corollary (6.2.28) is necessary.

3. Suppose that $A = [a_{ij}] \in M_n$, that λ is an eigenvalue of $|A| \equiv [|a_{ij}|]$, and that there is a vector $x = [x_i] \in \mathbf{R}^n$ with all $x_i > 0$ such that $|A|x = \lambda x$. Let $D = \text{diag}(x_1, x_2, \ldots, x_n)$. Show that every Geršgorin circle of $D^{-1}|A|D$

passes through λ. Draw a picture. What can you say about the absolute row sums of $D^{-1}AD$?

4. It will be proved in Chapter 8 that a square matrix with positive entries always has a positive eigenvalue and an associated positive eigenvector. Use this fact and the preceding problem to show that

$$\rho(A) \leq \rho(|A|)$$

whenever all entries of A are nonzero. Argue by continuity that the last requirement may be dropped, so

$$\rho(A) \leq \rho(|A|) \quad \text{for all} \quad A \in M_n$$

5. Use Corollary (6.2.28) to show that Cauchy's bound (5.6.40) on the zeroes of the polynomial

$$p(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_1 z + a_0, \qquad a_0 \neq 0$$

can be improved slightly to

$$|z| < \max\{|a_0|, |a_1|+1, |a_2|+1, \cdots, |a_{n-1}|+1\}$$

under the assumption that it is *not* the case that

$$|a_0| = |a_1|+1 = |a_2|+1 = \cdots = |a_{n-1}|+1$$

Hint: Show that the companion matrix $C(p)$ in (5.6.39) is irreducible if $a_0 \neq 0$. What improvements can be made in Montel's bound (5.6.41), Carmichael and Mason's bound (5.6.42), Montel's bound (5.6.43), and Kojima's bound (5.6.45)?

Further Reading. For a discussion of the Levy-Desplanques theorem and many references to the literature, see O. Taussky, "A Recurring Theorem on Determinants," *Amer. Math. Monthly* 56 (1949), 672–676.

6.3 Perturbation theorems

Let $D = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \in M_n$, let $E = [e_{ij}] \in M_n$, and consider the perturbed matrix $D+E$. By Theorem (6.1.1) the eigenvalues of $D+E$ are contained in the discs

$$\left\{z \in \mathbf{C} : |z - \lambda_i - e_{ii}| \leq R_i'(E) = \sum_{\substack{j=1 \\ j \neq i}}^{n} |e_{ij}|\right\}, \qquad i = 1, \ldots, n$$

which are contained in the discs

$$\left\{z \in \mathbf{C} : |z - \lambda_i| \leq R_i(E) = \sum_{j=1}^{n} |e_{ij}|\right\}, \qquad i = 1, \ldots, n$$

Thus, if $\hat{\lambda}$ is an eigenvalue of $D+E$, there is some eigenvalue λ_i of D such that $|\hat{\lambda} - \lambda_i| \leq \|E\|_\infty$. Unfortunately, this simple estimate does not extend to the general (nondiagonal) case, but we can use it to give a simple bound in the case in which the matrix is diagonalizable.

6.3.1 Observation. Let $A \in M_n$ be diagonalizable with $A = S\Lambda S^{-1}$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$. Let $E \in M_n$. If $\hat{\lambda}$ is an eigenvalue of $A+E$, then there is some eigenvalue λ_i of A for which

$$|\hat{\lambda} - \lambda_i| \leq \|S\|_\infty \|S^{-1}\|_\infty \|E\|_\infty = \kappa_\infty(S)\|E\|_\infty$$

where $\kappa_\infty(\cdot)$ denotes the condition number with respect to the matrix norm $\|\cdot\|_\infty$.

Proof: Since $A+E$ and $S^{-1}(A+E)S = \Lambda + S^{-1}ES$ have the same eigenvalues and since Λ is diagonal, the previous argument shows that there is some λ_i for which $|\hat{\lambda} - \lambda_i| \leq \|S^{-1}ES\|_\infty$. The stated inequality follows since $\|\cdot\|_\infty$ is a matrix norm. $\square$

By a slight change in technique, we can generalize this result to matrix norms other than the maximum row sum norm. The key hypothesis on the matrix norm is satisfied for all induced matrix norms that are induced by a monotone or absolute vector norm; see (5.6.37).

6.3.2 Theorem. Let $A \in M_n$ be diagonalizable with $A = S\Lambda S^{-1}$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$. Let $E \in M_n$ and let $\|\cdot\|$ be a matrix norm such that $\|D\| = \max_{1 \leq i \leq n}|d_i|$ for all diagonal matrices $D = \text{diag}(d_1, \ldots, d_n) \in M_n$. If $\hat{\lambda}$ is an eigenvalue of $A+E$, then there is some eigenvalue λ_i of A for which

$$|\hat{\lambda} - \lambda_i| \leq \|S\|\,\|S^{-1}\|\,\|E\| = \kappa(S)\|E\| \tag{6.3.3}$$

where $\kappa(\cdot)$ is the condition number with respect to the matrix norm $\|\cdot\|$.

Proof: As in the previous result, it suffices to consider the eigenvalues of $S^{-1}(A+E)S = \Lambda + S^{-1}ES$. If $\hat{\lambda}$ is an eigenvalue of $\Lambda + S^{-1}ES$, then $\hat{\lambda}I - \Lambda - S^{-1}ES$ is singular. If $\hat{\lambda}I - \Lambda$ is singular, then $\hat{\lambda} = \lambda_i$ for some i and the bound (6.3.3) is trivially satisfied. Suppose, however, that $\hat{\lambda}I - \Lambda$ is nonsingular. In this case, the matrix

$$(\hat{\lambda}I - \Lambda)^{-1}(\hat{\lambda}I - \Lambda - S^{-1}ES) = I - (\hat{\lambda}I - \Lambda)^{-1}S^{-1}ES$$

is singular, and hence by (5.6.16) it must be that $|||(\hat{\lambda}I-\Lambda)^{-1}S^{-1}ES|||\geq 1$. Thus, because of the assumption made about the behavior of the matrix norm $|||\cdot|||$ on diagonal matrices, we have

$$1\leq |||(\hat{\lambda}I-\Lambda)^{-1}S^{-1}ES|||\leq |||S^{-1}ES|||\,|||(\hat{\lambda}I-\Lambda)^{-1}|||$$

$$=|||S^{-1}ES|||\max_{1\leq i\leq n}|\hat{\lambda}-\lambda_i|^{-1}=\frac{|||S^{-1}ES|||}{\min_{1\leq i\leq n}|\hat{\lambda}-\lambda_i|}$$

and hence

$$\min_{1\leq i\leq n}|\hat{\lambda}-\lambda_i|\leq |||S^{-1}ES|||\leq |||S^{-1}|||\,|||S|||\,|||E|||=\kappa(S)|||E|||. \quad \square$$

Exercise. Show that the assumption of the theorem concerning the behavior of the matrix norm on diagonal matrices is satisfied for all of the following norms: $|||\cdot|||_2$, $|||\cdot|||_\infty$, $|||\cdot|||_1$. Give an example of at least one other matrix norm that satisfies the assumption.

Exercise. Give an example of a matrix norm that does not satisfy the assumption of the theorem.

Exercise. Show that $|||U|||_2=1$ for any unitary matrix U.

Although the condition number $\kappa(\cdot)$ arose previously in (5.8) in the context of error bounds for solutions of linear equations, we see that it now arises in (6.3.3) as an upper bound on the ratio of errors

$$\frac{|\hat{\lambda}-\lambda_i|}{|||E|||}\leq\kappa(S)$$

in computing eigenvalues of a diagonalizable matrix. If $\kappa(S)$ is small (near 1), then small perturbations in the data may perturb the eigenvalues, but changes in the eigenvalues will be bounded by a term of the same order as the changes in the data. If $\kappa(S)$ is very large, however, then small perturbations in the data may result in relatively large changes in the eigenvalues.

Unlike the situation in (5.8) for solutions of linear equations, it is not $\kappa(A)$ that is of importance here, but $\kappa(S)$, where $A=S\Lambda S^{-1}$ and S is a matrix whose columns are eigenvectors of A. The condition number with respect to the spectral norm has the geometrical interpretation that $\kappa(S)=\cot(\theta/2)$, where θ is the least angle between Sx and Sy as x and y range over all possible orthogonal nonzero vectors [See Example (7.4.26)]. Thus, independent of the condition number of A, if a pair of linearly independent eigenvectors of A is nearly parallel, then two columns

of S (say, columns p and q for $p\neq q$) may be nearly parallel, and hence the angle between Se_p and Se_q may be small even though the unit basis vectors e_p and e_q are orthogonal. In this event, the spectral condition number $\kappa(S)$ will be large, and the problem of determining the eigenvalues of A may be ill conditioned.

If S is unitary (or nearly unitary), however, then S will take pairs of orthogonal vectors into orthogonal (or nearly orthogonal) vectors and the spectral condition number of S will be small (equal to 1, in fact, if S is unitary). In this case, the problem of determining the eigenvalues of A must be well conditioned. Of course, a matrix can be (exactly) unitarily diagonalized if and only if it is normal, so (6.3.2) yields a perturbation theorem for the full class of normal (in particular, Hermitian or real symmetric) matrices, which is of the same simple form as our original observation about diagonal matrices. Normal matrices are perfectly conditioned with respect to eigenvalue computations.

6.3.4 **Corollary.** Let $A\in M_n$ be a normal matrix with eigenvalues $\lambda_1,\ldots,\lambda_n$, and let $E\in M_n$. If $\hat{\lambda}$ is an eigenvalue of $A+E$, then there is some eigenvalue λ_i of A for which $|\hat{\lambda}-\lambda_i|\leq |||E|||_2$.

Notice that neither the perturbation matrix E nor the perturbed matrix $A+E$ need be normal. Corollary (6.3.4) is most often applied in the case of a real symmetric matrix A.

Exercise. Provide the details for a proof of Corollary (6.3.4).

Exercise. Weyl's theorem (4.3.1) can be used to give a better bound than the bound in (6.3.4) if one knows that both A and E are Hermitian. If $A, E\in M_n$ are Hermitian, if $\lambda_1\leq\lambda_2\leq\cdots\leq\lambda_n$ are the ordered eigenvalues of A, if $\hat{\lambda}_1\leq\hat{\lambda}_2\leq\cdots\leq\hat{\lambda}_n$ are the ordered eigenvalues of $A+E$, and if $\lambda_1(E)\leq\cdots\leq\lambda_n(E)$ are the ordered eigenvalues of E, use the inequalities (4.3.2) to show that

$$\lambda_1(E)\leq\hat{\lambda}_k-\lambda_k\leq\lambda_n(E)\quad\text{for all}\quad k=1,2,\ldots,n$$

and that

$$|\hat{\lambda}_k-\lambda_k|\leq\rho(E)=|||E|||_2$$

Explain why this is a better bound than (6.3.4). What information does this give if all the eigenvalues of E are known to be nonnegative?

It is not uncommon in numerical applications to have a situation in which both the original matrix A and the perturbing matrix E are real

and symmetric. In this case, and in the more general situation in which both A and $A+E$ are normal, there is a comprehensive bound available on the perturbations to all the eigenvalues.

6.3.5 Theorem (Hoffman and Wielandt). Let $A, E \in M_n$, assume that A and $A+E$ are both normal, let $\{\lambda_1, \ldots, \lambda_n\}$ be the eigenvalues of A in some given order, and let $\{\hat{\lambda}_1, \ldots, \hat{\lambda}_n\}$ be the eigenvalues of $A+E$ in some order. Then there exists a permutation $\sigma(i)$ of the integers $1, 2, \ldots, n$ such that

$$\left[\sum_{i=1}^{n} |\hat{\lambda}_{\sigma(i)} - \lambda_i|^2\right]^{1/2} \leq \|E\|_2 \tag{6.3.6}$$

Proof: Let $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$, let $\hat{\Lambda} = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_n)$, let $V \in M_n$ be a unitary matrix such that $A = V\Lambda V^*$, and let $W \in M_n$ be a unitary matrix such that $A+E = W\hat{\Lambda}W^*$. Then because the Frobenius norm is unitarily invariant, we have

$$\begin{aligned}
\|E\|_2^2 &= \|(A+E) - A\|_2^2 \\
&= \|W\hat{\Lambda}W^* - V\Lambda V^*\|_2^2 \\
&= \|V^*W\hat{\Lambda}W^*V - \Lambda\|_2^2 \\
&= \|Z\hat{\Lambda}Z^* - \Lambda\|_2^2 \\
&= \text{tr}(Z\hat{\Lambda}Z^* - \Lambda)(Z\hat{\Lambda}Z^* - \Lambda)^* \\
&= \text{tr}(\hat{\Lambda}\hat{\Lambda}^* + \Lambda\Lambda^*) - \text{tr}(Z\hat{\Lambda}Z^*\Lambda^* + \Lambda Z\hat{\Lambda}^*Z^*) \\
&= \sum_{i=1}^{n} (|\hat{\lambda}_i|^2 + |\lambda_i|^2) - 2\,\text{Re}\,\text{tr}(Z\hat{\Lambda}Z^*\Lambda^*)
\end{aligned}$$

where we have set $Z \equiv V^*W$. This representation makes it clear that

$$\|E\|_2^2 \geq \sum_{i=1}^{n} (|\hat{\lambda}_i|^2 + |\lambda_i|^2) - 2\max\{\text{Re}\,\text{tr}(U\hat{\Lambda}U^*\Lambda^*) : U \text{ is unitary}\} \tag{6.3.7}$$

and we shall show that the exact value of this lower bound is the asserted bound (6.3.6). If $U \equiv [u_{ij}] \in M_n$, one computes easily that

$$\text{Re}\,\text{tr}(U\hat{\Lambda}U^*\Lambda^*) = \sum_{i,j=1}^{n} |u_{ij}|^2 \text{Re}(\bar{\lambda}_i\hat{\lambda}_j)$$

and we are interested in the maximum value of this expression as U ranges over the compact set of all n-by-n unitary matrices. If we set $c_{ij} \equiv |u_{ij}|^2$ and let $C \equiv [c_{ij}]$, the matrix $C \in M_n$ will be a matrix with nonnegative entries and all of its row and column sums will be exactly +1 (because $UU^* = U^*U = I$). Thus, C is a doubly stochastic matrix whenever U is unitary, and if we modify our extremum problem to admit all doubly stochastic matrices, we shall gain the advantage of having an extremum problem over a convex compact set whose structure is known. The maximum over this larger domain is of course potentially larger:

$$\begin{aligned}
&\max\{\text{Re}\,\text{tr}(U\hat{\Lambda}U^*\Lambda^*) : U \text{ is unitary}\} \\
&= \max\left\{\sum_{i,j=1}^{n} |u_{ij}|^2 \text{Re}(\bar{\lambda}_i\hat{\lambda}_j) : U \text{ is unitary}\right\} \\
&\leq \max\left\{\sum_{i,j=1}^{n} c_{ij}\, \text{Re}(\bar{\lambda}_i\hat{\lambda}_j) : C \text{ is doubly stochastic}\right\}
\end{aligned}$$

But the function to be maximized is a linear function on a compact convex set, so the maximum occurs at an extreme point of the convex set (see Appendix B and note that a linear function is a convex function). The extreme points of the set of doubly stochastic matrices are the permutation matrices by Birkhoff's theorem (8.7.1), and hence there is a permutation matrix $P \in M_n$ such that

$$\max\left\{\sum_{i,j=1}^{n} c_{ij}\, \text{Re}(\bar{\lambda}_i\hat{\lambda}_j) : C \text{ is doubly stochastic}\right\} = \text{Re}\,\text{tr}(P\hat{\Lambda}P^T\Lambda^*)$$

Since a permutation matrix is unitary, we also have

$$\max\{\text{Re}\,\text{tr}(U\hat{\Lambda}U^*\Lambda^*) : U \text{ is unitary}\} = \text{Re}\,\text{tr}(P\hat{\Lambda}P^T\Lambda^*)$$

If $Pe_i = e_{\sigma(i)}$ for $i = 1, 2, \ldots, n$, then

$$\text{Re}\,\text{tr}(P\hat{\Lambda}P^T\Lambda^*) = \sum_{i=1}^{n} \text{Re}(\hat{\lambda}_{\sigma(i)}\bar{\lambda}_i)$$

and (6.3.7) says that

$$\begin{aligned}
\|E\|_2^2 &\geq \sum_{i=1}^{n} [|\hat{\lambda}_{\sigma(i)}|^2 + |\lambda_i|^2 - 2\,\text{Re}(\hat{\lambda}_{\sigma(i)}\bar{\lambda}_i)] \\
&= \sum_{i=1}^{n} |\hat{\lambda}_{\sigma(i)} - \lambda_i|^2
\end{aligned}$$

□

Theorem (6.3.5) says that there is a strong global stability to the set of eigenvalues of a normal matrix, but it does not tell which arrangement of the eigenvalues will fulfill the stated inequality. Not every arrangement will do, and indeed there is at least one arrangement for which the inequality in (6.3.6) is reversed (see Problem 7 at the end of this section). But in the important special case of Hermitian matrices, the natural ordering of the eigenvalues will do.

6.3.8 Corollary. Let $A, E \in M_n$, assume that A is Hermitian and thal $A+E$ is normal, let $\{\lambda_1, \ldots, \lambda_n\}$ be the eigenvalues of A arranged ir increasing order $(\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n)$, and let $\{\hat{\lambda}_1, \ldots, \hat{\lambda}_n\}$ be the eigen- values of $A+E$, ordered so that $\operatorname{Re}\hat{\lambda}_1 \le \operatorname{Re}\hat{\lambda}_2 \le \cdots \le \operatorname{Re}\hat{\lambda}_n$. Then

$$\left[\sum_{i=1}^{n} |\hat{\lambda}_i - \lambda_i|^2\right]^{1/2} \le \|E\|_2$$

Proof: By the theorem, there is some permutation σ of the given orde (increasing real parts) for the eigenvalues of $A+E$ for which

$$\left[\sum_{i=1}^{n} |\hat{\lambda}_{\sigma(i)} - \lambda_i|^2\right]^{1/2} \le \|E\|_2 \tag{6.3.9}$$

If the eigenvalues of $A+E$ in the list $\hat{\lambda}_{\sigma(1)}, \ldots, \hat{\lambda}_{\sigma(n)}$ are already in increas- ing order of their real parts, there is nothing to prove. If not, there ar two successive eigenvalues in the list that are not ordered in this way, say

$$\operatorname{Re}\hat{\lambda}_{\sigma(k)} > \operatorname{Re}\hat{\lambda}_{\sigma(k+1)} \quad \text{for some } k \text{ such that} \quad 1 \le k < n$$

But since

$$|\hat{\lambda}_{\sigma(k)} - \lambda_k|^2 + |\hat{\lambda}_{\sigma(k+1)} - \lambda_{k+1}|^2 = |\hat{\lambda}_{\sigma(k+1)} - \lambda_k|^2 + |\hat{\lambda}_{\sigma(k)} - \lambda_{k+1}|^2 + 2(\lambda_k - \lambda_{k+1})(\operatorname{Re}\hat{\lambda}_{\sigma(k+1)} - \operatorname{Re}\hat{\lambda}_{\sigma(k)})$$

and since $\lambda_k - \lambda_{k+1} \le 0$ by assumption, we see that

$$|\hat{\lambda}_{\sigma(k)} - \lambda_k|^2 + |\hat{\lambda}_{\sigma(k+1)} - \lambda_{k+1}|^2 \ge |\hat{\lambda}_{\sigma(k+1)} - \lambda_k|^2 + |\hat{\lambda}_{\sigma(k)} - \lambda_{k+1}|^2$$

Thus, the two eigenvalues $\hat{\lambda}_{\sigma(k)}$ and $\hat{\lambda}_{\sigma(k+1)}$ can be interchanged withou increasing the sum of squared differences. By a finite sequence of suc interchanges, the list of eigenvalues $\hat{\lambda}_{\sigma(1)}, \ldots, \hat{\lambda}_{\sigma(n)}$ can be transforme into the list $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_n$ in which the real parts are increasing and th asserted bound holds. ☐

In practice, the most common application of this corollary is to the ca in which both A and $A+E$ are Hermitian, or even real and symmetri

Exercise. Show that if $A, B \in M_n$ are Hermitian and if their eigenvalu are both arranged in increasing or decreasing order, then

$$\left(\sum_{i=1}^{n} [\lambda_i(A) - \lambda_i(B)]^2\right)^{1/2} \le \|A - B\|_2$$

Exercise. Show that the result in Theorem (6.3.5) need not be true if n both of A and $B = A+E$ are normal. *Hint:* Let $A = \begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix}$, $B = \begin{bmatrix} -1 & - \\ 1 & \end{bmatrix}$ [illegible] and show that

$$\sum_{i=1}^{2} [\lambda_i(A) - \lambda_i(B)]^2 = 16$$

for any ordering of the eigenvalues.

If A is not diagonalizable, there are no bounds known that are as simple as those in Theorem (6.3.2). It is possible, however, to derive an explicit formula which expresses how the *algebraically simple* eigenvalues (algebraic multiplicity equal to 1) of a matrix vary when the entries are perturbed. We require first a lemma on nonorthogonality of the left and right eigenvectors associated with a simple eigenvalue.

6.3.10 Lemma. If $A \in M_n$, if λ is an algebraically simple eigenvalue of A, and if x and y are right and left eigenvectors, respectively, corresponding to the eigenvalue λ of A, then $y^*x \ne 0$.

Proof: If $Ax = \lambda x$, $x \ne 0$, then we may use the procedure employed in the proof of the Schur triangularization theorem (2.3.1) to construct a unitary matrix U whose first column is $x/\|x\|_2$ and for which

$$U^*AU = \left[\begin{array}{c|c} \lambda & * \\ \hline 0 & B \end{array}\right], \qquad B \in M_{n-1}$$

Since λ is a simple eigenvalue of A, it cannot be an eigenvalue of B. The unit basis vector e_1 is the λ eigenvector of U^*AU. Now consider

$$(U^*AU)^* = U^*A^*U = \left[\begin{array}{c|c} \bar{\lambda} & 0 \\ \hline * & B^* \end{array}\right]$$

and suppose that $U^*A^*Uz = \bar{\lambda}z$ with $z \ne 0$. If $z^* = [0 \,|\, \xi^*]$, then $\xi \ne 0$ and ξ is a $\bar{\lambda}$ eigenvector of B^*. But then λ would be an eigenvalue of B, which is excluded by assumption. We conclude that z cannot have a zero first component; that is, $z^*e_1 \ne 0$. But then $(Uz)^*(Ue_1) = z^*e_1 \ne 0$, and the vectors Uz and Ue_1 are left and right λ eigenvectors of A. Since the left and right λ eigenspaces of A are one-dimensional by assumption, we must have $y = \alpha Uz$ for some $\alpha \ne 0$. But $x = \|x\|_2 Ue_1$, so it must be that $y^*x \ne 0$. ☐

Exercise. Consider $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ and show that the lemma is false if "algebraically simple" is omitted from the hypotheses.

Now suppose that λ is an algebraically simple eigenvalue of A. Then A has a uniquely (up to a scalar factor α, $|\alpha| = 1$) determined normalized right λ eigenvector x and a uniquely determined left λ eigenvector y, which

is normalized by the relation $y^*x=1$. If we consider a differentiable parameterization $A(t)$ such that $A(0)=A$ [e.g., $A(t)=A+tE$ for a fixed perturbation matrix E], then for all sufficiently small t there is a uniquely determined simple eigenvalue $\lambda(t)$ of $A(t)$ such that $\lambda(0)=\lambda$. There is also a right $\lambda(t)$ eigenvector $x(t)$, which is uniquely (up to a factor α as before) determined by the condition $x^*(t)x(t)\equiv 1$, and a left $\lambda(t)$ eigenvector $y(t)$, which is uniquely determined by the condition $y^*(t)x(t)\equiv 1$.

If we differentiate this last normalization condition, we obtain the identity

$$y'^*(t)x(t)+y^*(t)x'(t)=0 \tag{6.3.11}$$

Since $A(t)x(t)=\lambda(t)x(t)$ for all small t, we also have the identity $y^*(t)A(t)x(t)=\lambda(t)y^*(t)x(t)=\lambda(t)$. If we differentiate this identity, we obtain

$$\lambda'(t)=y'^*(t)A(t)x(t)+y^*(t)A'(t)x(t)+y^*(t)A(t)x'(t)$$

But since $A(t)x(t)=\lambda(t)x(t)$ and $y^*(t)A(t)=\lambda(t)y^*(t)$, this becomes

$$\lambda'(t)=\lambda(t)\{y'^*(t)x(t)+y^*(t)x'(t)\}+y^*(t)A'(t)x(t)=y^*(t)A'(t)x(t)$$

We have used the identity (6.3.11). At $t=0$ this is just the identity $\lambda'(0)=y^*A'(0)x$, subject to the normalizations $x^*x=1$ and $y^*x=1$. If x and y are right and left λ eigenvectors that are not necessarily normalized in this way, we can replace x by $x/(x^*x)^{1/2}$ and y by $(x^*x)^{1/2}y/x^*y$ to obtain the general identity $\lambda'(0)y^*x=y^*A'(0)x$. We have proved the following result for a matrix A, which need not be diagonalizable.

6.3.12 Theorem. Let $A(t)\in M_n$ be differentiable at $t=0$. Assume that λ is an algebraically simple eigenvalue of $A(0)$ and that $\lambda(t)$ is an eigenvalue of $A(t)$, for small t, such that $\lambda(0)=\lambda$. Let x be a right λ eigenvector of A and let y be a left λ eigenvector of A. Then

$$\lambda'(0)=\frac{y^*A'(0)x}{y^*x}$$

Exercise. Let $A(t)=A+tE$ for a fixed perturbation matrix E and show (under the assumptions of the theorem) that

$$\frac{d\lambda}{dt}=\frac{y^*Ex}{y^*x} \quad \text{at} \quad t=0$$

Exercise. Under the assumptioins of the theorem, show that

$$\frac{\partial\lambda}{\partial a_{ij}}=\frac{\bar{y}_i x_j}{y^*x}$$

for any i,j. This formula shows how λ varies with respect to changes in any element of A. *Hint:* Let $E=E_{ij}$, the n-by-n matrix whose only nonzero entry is a one in the i,j position.

Exercise. Consider the matrix $A=\begin{bmatrix}1&1\\0&1+\epsilon\end{bmatrix}$ and the eigenvalue $\lambda=1$, which is simple if $\epsilon\neq 0$. Compute $\partial\lambda/\partial a_{ij}$ explicitly for all four pairs i,j. How do these variations behave as $\epsilon\to 0$? Conclude that an eigenvalue λ can be very sensitive to certain perturbations in A if x and y are nearly orthogonal.

In contrast to the situation for eigen*values*, the eigen*vectors* of even a diagonalizable matrix may suffer radical changes with only small perturbations in the entries of the matrix. For example, if $A=\begin{bmatrix}1&0\\0&1\end{bmatrix}$ and $E=\begin{bmatrix}\epsilon&\delta\\0&0\end{bmatrix}$ for $\epsilon,\delta\neq 0$, then the eigenvalues of $A+E$ are $\lambda=1$ and $1+\epsilon$, and the respective normalized eigenvectors are

$$\frac{1}{(\epsilon^2+\delta^2)^{1/2}}\begin{bmatrix}-\delta\\ \epsilon\end{bmatrix} \quad \text{and} \quad \begin{bmatrix}1\\0\end{bmatrix}$$

By choosing the ratio of ϵ to δ appropriately, the first eigenvector can be chosen to point in any direction whatsoever, for ϵ and δ both arbitrarily small.

If we set $\epsilon=0$, then the perturbed matrix $A+E=\begin{bmatrix}1&\delta\\0&1\end{bmatrix}$ has only one independent eigenvector for any $\delta\neq 0$, whereas A itself has two independent eigenvectors.

All our estimates so far have been a priori bounds on the perturbations induced in the eigenvalues; they do not involve the computed eigenvalues or eigenvectors or any quantity derived from them. Suppose that an "approximate eigenvector" $\hat{x}\neq 0$ and an "approximate eigenvalue" $\hat{\lambda}$ have been found somehow. It may not be the case that $A\hat{x}$ is exactly equal to $\hat{\lambda}\hat{x}$, but when A is diagonalizable we can use the *residual vector* $r=A\hat{x}-\hat{\lambda}\hat{x}$ to obtain an estimate of how close $\hat{\lambda}$ is to an eigenvalue.

Write $A=S\Lambda S^{-1}$, and suppose that $\hat{\lambda}$ is not exactly equal to any eigenvalue of A. Then

$$r=A\hat{x}-\hat{\lambda}\hat{x}=S(\Lambda-\hat{\lambda}I)S^{-1}\hat{x}$$

so that $\hat{x}=S(\Lambda-\hat{\lambda}I)^{-1}S^{-1}r$. Then

$$\begin{aligned}\|\hat{x}\| &= \|S(\Lambda-\hat{\lambda}I)^{-1}S^{-1}r\| \le \|\!|S(\Lambda-\hat{\lambda}I)^{-1}S^{-1}\|\!|\,\|r\| \\ &\le \|\!|S\|\!|\,\|\!|S^{-1}\|\!|\,\|\!|(\Lambda-\hat{\lambda}I)^{-1}\|\!|\,\|r\| = \kappa(S)\,\|\!|(\Lambda-\hat{\lambda}I)^{-1}\|\!|\,\|r\| \\ &= \kappa(S)\left(\min_{1\le i\le n}|\lambda_i-\hat{\lambda}|\right)^{-1}\|r\|\end{aligned}$$

so that

$$\|\hat{x}\| \min_{1\le i\le n} |\lambda_i-\hat{\lambda}| \le \kappa(S)\|r\|$$

Obviously, the latter inequality holds even when some $\lambda_i=\hat{\lambda}$. For this argument, we have assumed that

$$\left.\begin{array}{ll}\text{(a)} & \|\cdot\| \text{ is a vector norm on } \mathbf{C}^n;\\ \text{(b)} & \text{The matrix norm } \|\cdot\| \text{ on } M_n \text{ is compatible with } \|\cdot\|; \text{ and}\\ \text{(c)} & \|D\| = \max_{1\le i\le n} |d_i| \text{ whenever } D=\operatorname{diag}(d_1,\ldots,d_n)\in M_n;\end{array}\right\} \quad (6.3.13)$$

and the condition number $\kappa(S)$ is computed using the matrix norm $\|\cdot\|$. If A is normal, S may be taken to be unitary, and if we use the l_2 vector norm and the spectral matrix norm, we have $\kappa(S)=1$. The condition (c) is equivalent to requiring that the matrix norm $\|\cdot\|$ be induced by a monotone vector norm [Theorem (5.6.37)]. Thus, all of the conditions (6.3.13) are met if $\|\cdot\|$ is a monotone vector norm on $\mathbf{C}^n$ and $\|\cdot\|$ is the matrix norm on M_n that is induced by $\|\cdot\|$. We have proved a result about a posteriori bounds which is of the same type as Theorem (6.3.2) and Corollary (6.3.4).

6.3.14 **Theorem.** Let $A\in M_n$ be diagonalizable with $A=S\Lambda S^{-1}$ and $\Lambda=\operatorname{diag}(\lambda_1,\ldots,\lambda_n)$. Let the vector norm $\|\cdot\|$ on $\mathbf{C}^n$ and the matrix norm $\|\cdot\|$ on M_n satisfy the conditions (6.3.13), let $\hat{x}\in\mathbf{C}^n$ be a given nonzero vector, let $\hat{\lambda}$ be a given complex number, and let $r=A\hat{x}-\hat{\lambda}\hat{x}$. Then there is some eigenvalue λ_i of A for which

$$|\hat{\lambda}-\lambda_i| \le \|S\|\,\|S^{-1}\|\frac{\|r\|}{\|\hat{x}\|} = \kappa(S)\frac{\|r\|}{\|\hat{x}\|} \qquad (6.3.15)$$

If A is normal, then there is some eigenvalue λ_i of A for which

$$|\hat{\lambda}-\lambda_i| \le \frac{\|r\|_2}{\|\hat{x}\|_2} \qquad (6.3.16)$$

The latter result should be contrasted with the corresponding result for a posteriori bounds on the relative error in the solution to a system of linear equations. If the matrix of coefficients of a system of linear equations is ill-conditioned, (5.8.11) says that a small residual does not imply a small relative error in the solution. However, (6.3.16) says that if A is normal (in practice, A will usually be Hermitian or real symmetric), and if an approximate eigenvector-eigenvalue pair has a small residual, then the absolute error in the eigenvalue is guaranteed to be small; no condition number appears in the bound.

This pleasant result for the eigenvalues is not matched by a similarly pleasant result for the eigenvectors. Even for a real symmetric matrix, a small residual does not guarantee that the approximate eigenvector is close to an eigenvector. For example, consider $A=\begin{bmatrix}1&\epsilon\\ \epsilon&1\end{bmatrix}$ for $\epsilon>0$. If we take $\hat{\lambda}=1$ and $\hat{x}=[1,0]^T$, then the residual is $r=[0,\epsilon]^T$. The eigenvectors of A are $[1,1]^T$ and $[1,-1]^T$ for all $\epsilon>0$, and $\hat{x}$ is not approximately parallel to either of these two vectors no matter how small ϵ is.

Exercise. Show that the eigenvalues of A in the preceding example are $1+\epsilon$ and $1-\epsilon$, and verify the bound (6.3.16) in this case.

Problems

1. If λ, μ are eigenvalues of A and $\lambda\ne\mu$, show that any left eigenvector of A corresponding to μ is orthogonal to any right eigenvector of A corresponding to λ.

2. Use the preceding problem to give an alternate proof of Lemma (6.3.10) under the assumption that A has distinct eigenvalues.

3. Verify the lament expressed in the last sentence of the first paragraph of this section by considering

$$A_\epsilon=\begin{bmatrix}0&1\\ \epsilon&0\end{bmatrix}\in M_2, \qquad A_0=\begin{bmatrix}0&1\\ 0&0\end{bmatrix}$$

where $\epsilon\ge0$ is small. Show that A_ϵ is diagonalizable for $\epsilon>0$ and that the minimum distance between an eigenvalue of A_ϵ and any eigenvalue of A_0 is $\sqrt{\epsilon}$. Write $A_\epsilon=A_0+E$, and show that

$$\frac{|\hat{\lambda}-\lambda_i|}{\|E\|} \ge O(\epsilon^{-1/2})\to\infty \quad \text{as} \quad \epsilon\to0$$

Thus, no bound of the form $|\hat{\lambda}-\lambda_i|\le\|E\|$ can be correct in general. Now evaluate the bounds in Theorem (6.3.2) in this case and explain what is happening.

4. Consider the polynomial $p(x)=(x-x_0)^2$, which has a double root at x_0, that is, $p(x_0)=p'(x_0)=0$, but $p''(x_0)=0$. Show that for small $\epsilon>0$, $p(x)-\epsilon$ has two roots near x_0 of the form $x_0\pm\epsilon^{1/2}$. Thus, a change of order ϵ in the coefficients of a polynomial can perturb its zeroes by an amount of the order $\sqrt{\epsilon}$. For a polynomial, the ratio of the perturbation in a zero to a perturbation in the coefficients can be unbounded.

5. Consider the bound (6.3.4), which says that for a Hermitian (or, more generally, normal) matrix, the ratio of the perturbation in the eigenvalue to the perturbation in the matrix elements is bounded. Since the eigenvalues of the matrix are just the roots of the characteristic equation, explain how this pleasant situation could be consistent with the conclusion of Problem 4. The moral is that, as a practical matter, it is very unwise to compute the eigenvalues of a Hermitian (or any other) matrix by forming the characteristic polynomial and then computing its zeroes. This has the potential of turning an inherently well-conditioned problem into an ill-conditioned one!

6. Consider Givens's example of a real symmetric 2-by-2 matrix $A=I$ and a real symmetric perturbation

$$E(\epsilon)=\begin{bmatrix}\epsilon\cos(2/\epsilon) & \epsilon\sin(2/\epsilon)\\ \epsilon\sin(2/\epsilon) & -\epsilon\cos(2/\epsilon)\end{bmatrix},\qquad \epsilon>0$$

with $E(0)\equiv\lim_{\epsilon\to0}E(\epsilon)=0$. Show that the eigenvalues of $A+E(\epsilon)$ are $1+\epsilon$ and $1-\epsilon$, and that the respective (uniquely determined up to sign) normalized real eigenvectors of $A+E(\epsilon)$ are $[\cos(1/\epsilon),\ \sin(1/\epsilon)]^T$ and $[\sin(1/\epsilon),\ -\cos(1/\epsilon)]^T$ for $\epsilon>0$. Show that as $\epsilon\to0$, each eigenvector points in any given direction infinitely often. Thus, even if we restrict attention to real symmetric matrices, an individual eigenvector may vary rapidly if its eigenvalue is not well separated from others.

7. Use the argument of Theorem (6.3.5) to show that (under the hypotheses of the theorem) there is a permutation τ of the integers $1,2,\ldots,n$ such that

$$\left(\sum_{i=1}^{n}|\hat\lambda_{\tau(i)}-\lambda_i|^2\right)^{1/2}\geq\|E\|_2$$

Hint: Consider $\min\{\sum_{i,j=1}^{n}c_{ij}\operatorname{Re}(\hat\lambda_i\bar\lambda_j): C=[c_{ij}]$ is doubly stochastic$\}$.

8. Let $A\in M_n$ be a given normal matrix with eigenvalues $\{\lambda_i(A)\}$, let $r>0$ be given, and define

$$S(A,r)\equiv\{B\in M_n: B \text{ is normal and } \|B-A\|_2\leq r\}$$

Show that $\{\hat\lambda_1,\ldots,\hat\lambda_n\}$ is the set of eigenvalues of a matrix $B\in S(A,r)$ if and only if

$$\min\left\{\sum_{i=1}^{n}|\lambda_i(A)-\hat\lambda_{\sigma(i)}|^2: \sigma \text{ is a permutation of } 1,\ldots,n\right\}\leq r^2$$

This gives a complete characterization of the possible sets of eigenvalues of normal matrices in a neighborhood of a given normal matrix. *Hint:* Use Theorem (6.3.5) for necessity. For sufficiency, suppose that $A=U\Lambda U^*$ with $\Lambda=\operatorname{diag}(\lambda_1(A),\ldots,\lambda_n(A))$ and define $B\equiv U\hat\Lambda U^*$ with $\hat\Lambda=\operatorname{diag}(\hat\lambda_1,\ldots,\hat\lambda_n)$.

9. In the proof of Theorem (6.3.5) we used the fact that if $U=[u_{ij}]\in M_n$ is unitary, then $A\equiv[|u_{ij}|^2]$ is doubly stochastic. Show that not every doubly stochastic matrix arises in this way from a unitary matrix. *Hint:* Consider the example

$$\frac{1}{2}\begin{bmatrix}1&1&0\\1&0&1\\0&1&1\end{bmatrix}$$

10. Suppose $A\in M_3$ is a given Hermitian matrix, and suppose one has found somehow a unitary matrix U such that

$$UAU^*=\begin{bmatrix}3.05&-.06&.02\\-.06&-6.91&.07\\.02&.07&8.44\end{bmatrix}$$

Give the best estimate you can for the eigenvalues of A.

11. There is no hope of having a bound of the type (6.3.4) for nonnormal matrices. Consider $A,E\in M_n$, where

$$A=\begin{bmatrix}0&a&&&\mathbf{0}\\&0&\ddots&&\\&&\ddots&\ddots&\\&&&0&\\\mathbf{0}&&&&0\end{bmatrix},$$

$$E=\begin{bmatrix}0&0&&&\mathbf{0}\\ \vdots&&\epsilon&&\\ &&&\epsilon\ddots&\ddots\\ 0&\mathbf{0}&&&\epsilon\\ \epsilon&0&\cdots&&0\end{bmatrix},\qquad a,\epsilon\geq0$$

Show that all the eigenvalues of A are 0, and the eigenvalues of $A+E$ are the n distinct values for $\sqrt[n]{a\epsilon^{n-1}}$. Whatever the value of $\epsilon>0$, all the eigenvalues of $A+E$ can be made arbitrarily large by a suitable choice of a. How is the situation different when A is normal?

Further Readings. The original version of Theorem (6.3.5) is in A. J. Hoffman and H. Wielandt, "The Variation of the Spectrum of a Normal

Matrix," *Duke Math. J.* 20 (1953), 37–39. An elementary proof of this result in the real symmetric case is in [Wil], pp. 104–109.

6.4 Other inclusion regions

We have discussed the Geršgorin discs in some detail. They are a particular class of easily computed regions in the plane that are guaranteed to include the eigenvalues of a given matrix. Many authors, perhaps attracted by the geometrical elegance of the Geršgorin theory, have generalized the ideas and methods of this theory to obtain other types of inclusion regions. We discuss a few of these to give the flavor of what has been done.

The first result gives an eigenvalue inclusion region that is a union of discs, like the Geršgorin region, but the radii of the discs depend on both the deleted row and column sums. The separate row and column sum version of Geršgorin's theorem are obtained as limits in this result, due to Ostrowski, which may therefore be viewed as giving a continuum of inclusion regions that interpolate between (6.1.2) and (6.1.4).

6.4.1 **Theorem** (Ostrowski). Let $A=[a_{ij}]\in M_n$, let $\alpha\in[0,1]$ be given, and let R_i' and C_i' denote the deleted row and column sums of A, respectively:

$$R_i'=\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}| \tag{6.4.2}$$

$$C_i'=\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ji}| \tag{6.4.3}$$

Then all the eigenvalues of A are located in the union of n discs

$$\bigcup_{i=1}^{n}\{z\in\mathbf{C}: |z-a_{ii}|\le R_i'^{\alpha}C_i'^{1-\alpha}\} \tag{6.4.4}$$

Proof: We may assume that $0<\alpha<1$, as the cases $\alpha=0$ and $\alpha=1$ (Geršgorin's theorems for column and row sums, respectively) can be obtained by taking limits. Furthermore, we may assume that all $R_i'>0$, because we may perturb A by inserting a small nonzero entry into any row in which $R_i'=0$; the resulting matrix has an inclusion region (6.4.4) that is larger than the region for A, and the result follows in the limit as the perturbation goes to zero.

Now suppose that $Ax=\lambda x$ with $x=[x_i]\neq 0$. Then for each $i=1,2,\ldots,n$ we have

$$|\lambda-a_{ii}|\,|x_i|=\left|\sum_{\substack{j=1\\ j\neq i}}^{n}a_{ij}x_j\right|\le\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|\,|x_j|=\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|^{\alpha}\{|a_{ij}|^{1-\alpha}|x_j|\}$$

$$\le\left[\sum_{\substack{j=1\\ j\neq i}}^{n}\{|a_{ij}|^{\alpha}\}^{1/\alpha}\right]^{\alpha}\left[\sum_{\substack{j=1\\ j\neq i}}^{n}\{|a_{ij}|^{1-\alpha}|x_j|\}^{1/(1-\alpha)}\right]^{1-\alpha} \tag{6.4.5a}$$

$$=R_i'^{\alpha}\left[\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|\,|x_j|^{1/(1-\alpha)}\right]^{1-\alpha}$$

which, since $R_i'>0$, is equivalent to

$$\frac{|\lambda-a_{ii}|}{R_i'^{\alpha}}|x_i|\le\left[\sum_{j=1}^{n}|a_{ij}|\,|x_j|^{1/(1-\alpha)}\right]^{1-\alpha}$$

and hence

$$\left[\frac{|\lambda-a_{ii}|}{R_i'^{\alpha}}\right]^{1/(1-\alpha)}|x_i|^{1/(1-\alpha)}\le\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|\,|x_j|^{1/(1-\alpha)} \tag{6.4.5b}$$

The inequality employed in (6.4.5a) is Hölder's inequality (Appendix B) with $p=1/\alpha$ and $q=p/(p-1)=1/(1-\alpha)$. Now sum (6.4.5b) on i to get

$$\sum_{i=1}^{n}\left[\frac{|\lambda-a_{ii}|}{R_i'^{\alpha}}\right]^{1/(1-\alpha)}|x_i|^{1/(1-\alpha)}\le\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|\,|x_j|^{1/(1-\alpha)}$$

$$=\sum_{j=1}^{n}C_j'|x_j|^{1/(1-\alpha)} \tag{6.4.6}$$

If

$$\left[\frac{|\lambda-a_{ii}|}{R_i'^{\alpha}}\right]^{1/(1-\alpha)}>C_i'$$

for every i such that $x_i\neq 0$, then (6.4.6) could not be correct. Thus, we conclude that

$$\left[\frac{|\lambda-a_{ii}|}{R_i'^{\alpha}}\right]^{1/(1-\alpha)}\le C_i'$$

for at least one value of i such that $x_i\neq 0$, and hence

$$|\lambda-a_{ii}|\le R_i'^{\alpha}C_i'^{1-\alpha}$$ □

Exercise. Consider $A=\begin{bmatrix}1&4\\1&6\end{bmatrix}$ and compare the Geršgorin row and column eigenvalue inclusion regions with Ostrowsi's region for $\alpha=\frac{1}{2}$. What estimate does Ostrowski's theorem give for the spectral radius of A and how does it compare with the Geršgorin estimates (6.1.5)?

Exercise. What is the Ostrowski version of Corollary (6.1.6)?

The next result, due to Brauer, is also a generalization of Geršgorin's theorem, but now the rows are taken two at a time; the geometrical regions are no longer discs but sets known as *ovals of Cassini*. The proof obviously parallels the proof of Geršgorin's theorem in that one selects not just the largest modulus component of the eigenvector, but the two largest modulus components.

6.4.7 **Theorem** (Brauer). Let $A=[a_{ij}]\in M_n$. All the eigenvalues of A are located in the union of $n(n-1)/2$ ovals of Cassini

$$\bigcup_{\substack{i,j=1\\ i\neq j}}^{n} \{z\in \mathbf{C}\colon |z-a_{ii}|\,|z-a_{jj}|\le R_i'R_j'\} \tag{6.4.8}$$

Proof: Let λ be an eigenvalue of A, and suppose that $Ax=\lambda x$ with $x=[x_i]\neq 0$. There is an element of x that has largest absolute value, say x_p, so $|x_p|\ge |x_i|$ for all $i=1,\dots,n$ and $x_p\neq 0$. If all the other entries of x are zero, then the assumption that $Ax=\lambda x$ means that $a_{pp}=\lambda$. Since all the diagonal entries of A are included in the region (6.4.8), the eigenvalue λ is in this region whenever its associated eigenvector has only one nonzero entry.

Now suppose that there are at least two nonzero entries of the eigenvector x, and let x_q be the component with second largest absolute value; that is, $|x_p|\ge|x_q|\ge|x_i|$ for all $i=1,\dots,n$, $i\neq p$, and $x_p\neq 0\neq x_q$. Then $Ax=\lambda x$ means that

$$x_p(\lambda-a_{pp})=\sum_{\substack{j=1\\ j\neq p}}^{n} a_{pj}x_j$$

which implies that

$$|x_p|\,|\lambda-a_{pp}|=\left|\sum_{\substack{j=1\\ j\neq p}}^{n} a_{pj}x_j\right|\le \sum_{\substack{j=1\\ j\neq p}}^{n} |a_{pj}|\,|x_j|\le \sum_{\substack{j=1\\ j\neq p}}^{n} |a_{pj}|\,|x_q|=R_p'|x_q|$$

or

$$|\lambda-a_{pp}|\le R_p'\frac{|x_q|}{|x_p|} \tag{6.4.9}$$

But we also have

$$x_q(\lambda-a_{qq})=\sum_{\substack{j=1\\ j\neq q}}^{n} a_{qj}x_j$$

which implies that

$$|x_q|\,|\lambda-a_{qq}|=\left|\sum_{\substack{j=1\\ j\neq q}}^{n} a_{qj}x_j\right|\le \sum_{\substack{j=1\\ j\neq q}}^{n} |a_{qj}|\,|x_j|\le \sum_{\substack{j=1\\ j\neq q}}^{n} |a_{qj}|\,|x_p|=R_q'|x_p|$$

or

$$|\lambda-a_{qq}|\le R_q'\frac{|x_p|}{|x_q|} \tag{6.4.10}$$

Taking the product of (6.4.9) and (6.4.10) permits us to eliminate the unknown ratios of components of x and obtain

$$|\lambda-a_{pp}|\,|\lambda-a_{qq}|\le R_p'\frac{|x_q|}{|x_p|}R_q'\frac{|x_p|}{|x_q|}=R_p'R_q'$$

Thus, the eigenvalue λ lies in the region (6.4.8). □

Exercise. What is the column sum version of Brauer's theorem?

Any theorem about eigenvalue inclusion regions implies (and, indeed, is implied by) a related theorem about invertibility. One just uses the inclusion result to set conditions that prohibit $z=0$ from being in the region.

6.4.11 **Corollary.** If $A=[a_{ij}]\in M_n$, then either of the following conditions is sufficient that A be invertible:

(a) For some $\alpha\in[0,1]$, $|a_{ii}|>R_i'^{\alpha}C_i'^{1-\alpha}$ for all $i=1,\dots,n$ (Ostrowskii)

(b) $|a_{ii}|\,|a_{jj}|>R_i'R_j'$ for all $i,j=1,\dots,n$, $i\neq j$ (Brauer)

Exercise. Use (6.4.1) and (6.4.7) to prove (6.4.11).

Brauer's theorem involves products of rows taken two at a time. An attractive possibility for further generalization is suggested by the idea of taking the rows three or more at a time, and considering, for each $m=1,\dots,n$, the union of sets of the form

$$\left\{z\in\mathbf{C}\colon \prod_{k=1}^{m}|z-a_{i_ki_k}|\le\prod_{k=1}^{m}R_{i_k}'\right\},\qquad A=[a_{ij}]\in M_n \tag{6.4.12}$$

For each m, there are $\binom{n}{m}$ sets of this form; $m=1$ gives the n Geršgorin discs and $m=2$ gives Brauer's $n(n-1)/2$ ovals of Cassini. Unfortunately, for $m\ge 3$ the sets (6.4.12) need not be be eigenvalue inclusion regions at all, as shown by the example

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6.4.13}$$

The sets (6.4.12) for $m=3$ and $m=4$ all collapse to the point $z=1$.

Exercise. Show that the eigenvalues of the matrix in (6.4.13) are $\lambda=0$, 1, 1, and 2. Sketch the sets (6.4.12) for $m=1$, $m=2$, and $m=3,4$. Show that the same phenomenon occurs for all $m\geq 3$ by considering

$$A = \begin{bmatrix} J & 0 \\ 0 & I_n \end{bmatrix} \in M_{n+2} \tag{6.4.14}$$

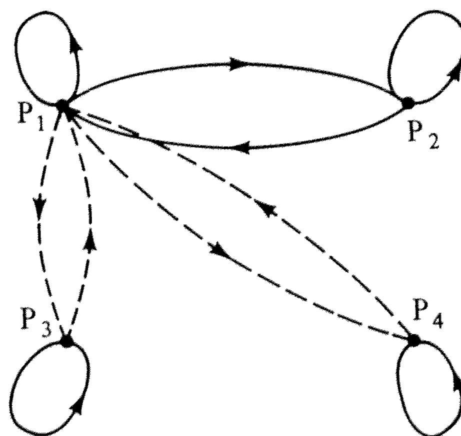
where $J = \left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right] \in M_2$ and $I_n \in M_n$ is the n-by-n identity matrix.

Although this example eliminates the most obvious generalization of Brauer's theorem, it suggests what may be wrong and how to deal with it. The problem with the region (6.4.12) is that it admits too many products, some of which may be zero because of zero deleted row sums. Of course, this can't happen if the matrix A is irreducible; all $R_i'>0$ in this case.

However, even if A is irreducible, the region (6.4.12) may not be an eigenvalue inclusion region for A; it may still admit too many products. Consider the perturbation of (6.4.13) given by

$$A_\epsilon = \begin{bmatrix} 1 & 1 & \epsilon & \epsilon \\ 1 & 1 & 0 & 0 \\ \epsilon & 0 & 1 & 0 \\ \epsilon & 0 & 0 & 1 \end{bmatrix}, \qquad 1>\epsilon\geq 0 \tag{6.4.15}$$

The directed graph $\Gamma(A_\epsilon)$ of A_ϵ is



where the dashed arcs disappear when $\epsilon=0$. If $\epsilon\neq 0$, $\Gamma(A_\epsilon)$ is strongly connected and A_ϵ is irreducible. We find that

$$R_1'=1+2\epsilon, \qquad R_2'=1, \qquad R_3'=\epsilon, \qquad R_4'=\epsilon$$

and A_ϵ has eigenvalues

$$\lambda_\epsilon = 1, 1, 1+(1+2\epsilon^2)^{1/2} \quad \text{and} \quad 1-(1+2\epsilon^2)^{1/2}$$

Exercise. Verify the above calculations for A_ϵ.

Since any product of three or more of the R_i' terms contains at least one factor of ϵ, the sets (6.4.12) cannot be eigenvalue inclusion regions for either $m=3$ or $m=4$ when ϵ is small and positive.

Exercise. Show that this conclusion holds for all $m\geq 3$ by considering a perturbation of (6.4.14) of the same form as (6.4.15).

What intrinsic property of (6.4.13) and (6.4.15) indicates that $m=1$ and $m=2$ are acceptable in (6.4.12), but that $m=3$ and $m=4$ are not? Richard Brualdi noticed that the directed graphs in each case do not contain any cycles of length 3 or 4, but do contain cycles of length 1 and 2. This turns out to be the key observation in obtaining a correct generalization of Brauer's theorem.

Recall that a directed graph Γ is *strongly connected* if and only if from each node there is a directed path in Γ to *any other* node (and back).

We say that Γ is *weakly connected* if and only if from each node there is a directed path to *some other* node and then back. This is equivalent to the assertion that *each node in* Γ *belongs to some nontrivial cycle;* a trivial cycle (or loop) is a directed path of length 1 that begins and ends at the same node.

In terms of matrices, we know that $\Gamma(A)$ is strongly connected if and only if A is irreducible. We say that A is *weakly irreducible* if and only if $\Gamma(A)$ is weakly connected. Weak irreducibility does not seem to have as striking a characterization as irreducibility [in terms of something like the permutation similarity (6.2.21b)], but in terms of the zero–nonzero structure of the entries of A it is clear that A is weakly irreducible if and only if for each $i=1,\ldots,n$ the ith row of A has at least one nonzero off-diagonal entry a_{ij_i} such that there is a sequence $a_{k_1k_2}, a_{k_2k_3}, \ldots, a_{k_{m-1}k_m}$ of nonzero entries of A for which $k_1=j_i$ and $k_m=i$. This cumbersome condition is about half of the requirement (6.2.7) that A have property *SC*, and it is perhaps more conveniently stated for computational purposes in a form analogous to Theorem (6.2.23).

6.4.16 Lemma. If $A \in M_n$, then A is weakly irreducible if and only if either of the matrices

(a) $B = [I + |A|]^{n-1}$ or
(b) $B = [I + M(A)]^{n-1}$

has the property that for each $i = 1, \ldots, n$ there is at least one nonzero off-diagonal entry b_{ij} in the ith row ($j \neq i$) such that b_{ji} is nonzero as well.

Exercise. Prove Lemma (6.4.16). *Hint:* Use the ideas in (6.2.19).

Exercise. Suppose $A \in M_n$ and let $B \in M_n$ be defined by either (a) or (b) in (6.4.16). Show that A is weakly irreducible if and only if $\Gamma(B)$ has the property that every node belongs to a cycle of length 2. What is the corresponding property for irreducible? Which property is weaker? Recall that a cycle is *simple* by definition; only the initial (which is the same as the final) node can appear in the list of nodes more than once.

Exercise. If $A \in M_n$ is weakly irreducible, show that all $R_i' > 0$ and all $C_i' > 0$.

A *preorder* on a set S is a relation R defined between *all* pairs of points of S such that for any pair of elements $s, t \in S$, either sRt or tRs or both. A preorder must also be reflexive (sRs for every $s \in S$) and transitive (if sRt and tRu, then sRu). A preorder might not be symmetric (sRt whenever tRs), and it could be that sRt and tRs without $s = t$. A point z in a subset S_0 of S is said to be *a maximal element* of S_0 if sRz for all $s \in S_0$.

Exercise. Let S be any nonempty set of complex numbers. Show that the relation between pairs of complex numbers $z, w \in S$ defined by

$$zRw \quad \text{if and only if} \quad |z| \leq |w|$$

is a preorder on $\mathbf{C}$.

6.4.17 Lemma. Let S be a nonempty finite set on which there is defined a preorder. Then S contains at least one maximal element.

Proof: Arrange the elements in any order $s_1, \ldots, s_k$. Set $s \equiv s_1$. If $s_2 R s$, then leave s alone, but if not, then set $s \equiv s_2$. Continue this process with the rest of the elements. The final value of s is a maximal element. □

If Γ is a directed graph and if P_i is a node of Γ, we define $\Gamma_{\text{out}}(P_i)$ to be the set of nodes different from P_i that can be reached from P_i by some directed path of length 1. Notice that if Γ is weakly connected, then $\Gamma_{\text{out}}(P_i)$ is nonempty for every node $P_i \in \Gamma$.

Let us denote by $C(A)$ the set of nontrivial cycles γ in the directed graph $\Gamma(A)$. A nontrivial cycle is one that contains at least two distinct nodes; that is, it is a (simple directed) cycle that is not a loop. For the matrix (6.4.13), $C(A)$ consists of the single cycle $\gamma = P_1P_2, P_2P_1$, while for the matrix (6.4.15) there are three separate nontrivial cycles, all of length 2.

6.4.18 Theorem (Brualdi). If $A = [a_{ij}] \in M_n$ is weakly irreducible, then every eigenvalue of A is contained in the region

$$\bigcup_{\gamma \in C(A)} \left\{ z \in \mathbf{C}: \prod_{P_i \in \gamma} |z - a_{ii}| \leq \prod_{P_i \in \gamma} R_i' \right\} \tag{6.4.19}$$

The notation means that if $\gamma = P_{i_1}P_{i_2}, \ldots, P_{i_k}P_{i_{k+1}}$ is a nontrivial cycle with $P_{i_{k+1}} = P_{i_1}$, then each of the products in (6.4.19) contains exactly k terms, and the index i takes on the k values $i_1, i_2, \ldots, i_k$.

Proof: Suppose λ is an eigenvalue of A and suppose $\lambda = a_{ii}$ for some main diagonal entry of A. Then λ is obviously in the region (6.4.19). In fact, all $R_i' > 0$ because A is weakly irreducible, so in this case λ lies in the interior of the region (6.4.19). If each of the eigenvalues of A is equal to some main diagonal entry of A, then all the eigenvalues of A lie in the interior of the region (6.4.19) and we are done.

For the rest of the argument, we suppose that λ is an eigenvalue of A such that $\lambda \neq a_{ii}$ for all $i = 1, \ldots, n$. Let $Ax = \lambda x$ for some nonzero $x = [x_i] \in \mathbf{C}^n$. Define a preorder R on the nodes of Γ by

$$P_i R P_j \quad \text{if and only if} \quad |x_i| \leq |x_j| \tag{6.4.20}$$

We shall show that there exists a cycle γ' in $\Gamma(A)$ with the following three properties:

(6.4.21)

(a) $\gamma' = P_{i_1}P_{i_2}, P_{i_2}P_{i_3}, \ldots, P_{i_k}P_{i_{k+1}}$ is a nontrivial (simple directed) cycle with $k \geq 2$ and $P_{i_{k+1}} = P_{i_1}$.

(b) For each $j = 1, \ldots, k$, the node $P_{i_{j+1}}$ is a maximal node in $\Gamma_{\text{out}}(P_{i_j})$; that is, $|x_{i_{j+1}}| \geq |x_m|$ for all m such that $P_m \in \Gamma_{\text{out}}(P_{i_j})$.

(c) All $x_{i_j} \neq 0$, $j = 1, \ldots, k$.

If γ' is a cycle that satisfies the conditions (6.4.21), then $Ax=\lambda x$ implies that for any $j=1,\ldots,k$ we have

$$(\lambda-a_{i_ji_j})x_{i_j}=\sum_{\substack{m=1\\m\neq i_j}}^{n} a_{i_jm}x_m=\sum_{P_m\in\Gamma_{\text{out}}(P_{i_j})} a_{i_jm}x_m$$

and hence

$$|\lambda-a_{i_ji_j}|\,|x_{i_j}|=\left|\sum_{P_m\in\Gamma_{\text{out}}(P_{i_j})} a_{i_jm}x_m\right|\leq\sum_{P_m\in\Gamma_{\text{out}}(P_{i_j})}|a_{i_jm}|\,|x_m| \tag{6.4.22}$$

$$\leq\sum_{P_m\in\Gamma_{\text{out}}(P_{i_j})}|a_{i_jm}|\,|x_{i_{j+1}}| \tag{6.4.22a}$$

$$=R'_{i_j}|x_{i_{j+1}}|$$

If we now take the product of the inequalities (6.4.22) over all the nodes in γ', we obtain

$$\prod_{j=1}^{k}|\lambda-a_{i_ji_j}|\,|x_{i_j}|\leq\prod_{j=1}^{k}R'_{i_j}|x_{i_{j+1}}| \tag{6.4.23}$$

But

$$\prod_{j=1}^{k}|\lambda-a_{i_ji_j}|=\prod_{P_i\in\gamma'}|\lambda-a_{ii}|\quad\text{and}\quad\prod_{j=1}^{k}R'_{i_j}=\prod_{P_i\in\gamma'}R'_i$$

and since $P_{i_{k+1}}=P_{i_1}$, we also have $x_{i_{k+1}}=x_{i_1}$. Therefore,

$$\prod_{j=1}^{k}|x_{i_j}|=\prod_{j=1}^{k}|x_{i_{j+1}}|\neq 0 \tag{6.4.24}$$

Thus, dividing (6.4.23) by (6.4.24), we obtain

$$\prod_{P_i\in\gamma'}|\lambda-a_{ii}|\leq\prod_{P_i\in\gamma'}R'_i \tag{6.4.25}$$

Since γ' is a nontrivial cycle in $\Gamma(A)$, the eigenvalue λ must therefore lie in the region (6.4.19).

We must now show that there is a cycle γ' that satisfies the conditions (6.4.21). Let i be any index for which $x_i\neq 0$. Then from the identity

$$(\lambda-a_{ii})x_i=\sum_{\substack{j=1\\j\neq i}}^{n}a_{ij}x_j=\sum_{P_j\in\Gamma_{\text{out}}(P_i)}a_{ij}x_j$$

and the fact that $x_i\neq 0$ and $\lambda-a_{ii}\neq 0$, we see that the left-hand side is nonzero and hence among the nodes of $\Gamma_{\text{out}}(P_i)$ [those P_j such that $a_{ij}\neq 0$ and $j\neq i$; $\Gamma_{\text{out}}(P_i)$ is nonempty because $\Gamma(A)$ is weakly connected] there

must be at least one for which the corresponding eigenvector component x_j is nonzero. Let $P_{i_1}\equiv P_i$, and let P_{i_2} be a maximal node among the nodes in $\Gamma_{\text{out}}(P_{i_1})$, that is, $|x_{i_2}|\geq|x_m|$ for all m such that $P_m\in\Gamma_{\text{out}}(P_{i_1})$. We are guaranteed that $x_{i_2}\neq 0$.

Suppose that the preceding construction has produced a directed path $P_{i_1}P_{i_2}, P_{i_2}P_{i_3},\ldots,P_{i_{j-1}}P_{i_j}$ of length $j-1$ that satisfies conditions (b) and (c) of (6.4.21); we have just done this for $j=2$. Then

$$(\lambda-a_{i_ji_j})x_{i_j}=\sum_{P_m\in\Gamma_{\text{out}}(P_{i_j})}a_{i_jm}x_m$$

and the left-hand side is nonzero, so there must be at least one node in $\Gamma_{\text{out}}(P_{i_j})$ [nonempty because $\Gamma(A)$ is weakly connected] for which the corresponding eigenvector component is nonzero. Thus, if we choose $P_{i_{j+1}}$ to be a maximal node in $\Gamma_{\text{out}}(P_{i_j})$, we are guaranteed that $x_{i_{j+1}}\neq 0$.

Because there are only finitely many nodes in $\Gamma(A)$, this construction for $j=2,3,\ldots$ must eventually produce a first maximal node $P_{i_q}\in\Gamma_{\text{out}}(P_{i_{q-1}})$ which was produced as a node P_{i_p} at some previous step $(2\leq p+1<q)$. Then $\gamma'=P_{i_p}P_{i_{p+1}}, P_{i_{p+1}}P_{i_{p+2}},\ldots,P_{i_{q-1}}P_{i_q}$ is a cycle in $\Gamma(A)$ which satisfies all three conditions in (6.4.21). □

Brualdi's theorem has a sharper form when A is actually irreducible; it is the generalized Brauer (6.4.7) version of Theorem (6.2.26).

6.4.26 Theorem (Brualdi). Let $A=[a_{ij}]\in M_n$ be irreducible. A boundary point λ of the region (6.4.19) can be an eigenvalue of A only if the boundary of each set

$$\left\{z\in\mathbf{C}:\prod_{P_i\in\gamma}|z-a_{ii}|\leq\prod_{P_i\in\gamma}R'_i\right\} \tag{6.4.27}$$

passes through λ for every nontrivial cycle $\gamma\in C(A)$.

Proof: Since all $R'_i>0$, if $\lambda=a_{ii}$ for any $i=1,2,\ldots,n$, then λ could not be on the boundary of the region (6.4.27). Thus, we may assume that $\lambda\neq a_{ii}$ for all $i=1,\ldots,n$ and we may continue the argument used in Brualdi's theorem (6.4.18) with the same notation, but with the additional assumption that λ is an eigenvalue of A that lies on the boundary of the region (6.4.19). Just as in the proof of Lemma (6.2.3), λ must satisfy the inequality

$$\prod_{P_i\in\gamma}|\lambda-a_{ii}|\geq\prod_{P_i\in\gamma}R'_i$$

for all nontrivial cycles $\gamma\in C(A)$ [with equality for at least one $\gamma\in C(A)$]. Comparing this inequality with (6.4.25), we see that

$$\prod_{P_i \in \gamma'} |\lambda - a_{ii}| = \prod_{P_i \in \gamma'} R_i' \tag{6.4.28}$$

for the particular cycle γ' constructed in the proof of (6.4.18). Thus, the inequality in (6.4.23) must be an equality, as must both of the inequalities in (6.4.22) for all $j = 1, 2, \ldots, k$. In particular, the inequality (6.4.22a) must be an equality, and hence for each $P_{i_j} \in \gamma'$ and for all m such that $P_m \in \Gamma_{\text{out}}(P_{i_j})$, $|x_m| = |x_{i_{j+1}}| = c_{i_{j+1}} =$ constant. Notice that this conclusion follows for any cycle that satisfies the conditions (6.4.21).

Now define the set

$$K \equiv \{P_i \in \Gamma(A): |x_m| = c_i = \text{constant for all } m \text{ such that } P_m \in \Gamma_{\text{out}}(P_i)\}$$

We know that K is not empty because all the nodes of γ' are in K. We would like to show that all the nodes of $\Gamma(A)$ are in K.

Suppose there is a node P_q of $\Gamma(A)$ that is not in K. Because $\Gamma(A)$ is strongly connected, there is at least one directed path in $\Gamma(A)$ from each node of K to this external node P_q. If we select from all such directed paths a path with shortest length, then its first arc must be from a node in K to a node P_f that is not in K. If we use the same preorder on the nodes of $\Gamma(A)$ that we used in the proof of Theorem (6.4.18), then we may employ the same construction used in the proof of Theorem (6.4.18); start with the node $P_f \equiv P_{j_1}$, select a maximal node $P_{j_2} \in \Gamma_{\text{out}}(P_{j_1})$, select a maximal node $P_{j_3} \in \Gamma_{\text{out}}(P_{j_2})$, and so on. At each step, $\Gamma_{\text{out}}(P_{j_i})$ is nonempty because $\Gamma(A)$ is weakly (even strongly) connected, and the maximal node satisfies condition (c) of (6.4.21) for the same reason as before.

If, at some step of this construction, we have a choice between selecting a maximal node which is in K or not in K, we shall always choose one that is *not* in K. If at any step, all the maximal nodes from which we may choose are in K, choose any one of them and then follow a directed path of shortest length (necessarily in K) to a first node that is not in K and resume selecting maximal nodes as before. By definition of K, *any* directed path in K will have the property that each node is a maximal node in Γ_{out} of its predecessor node [condition (b) of (6.4.21)]. Because the complement of K has only finitely many nodes, this construction must ultimately produce a first maximal node in the complement of K that was produced as a node at some previous step. The directed path between the first and second occurrences of this node in the construction will be a nontrivial directed cycle, which may not be simple because of the way we have forced the path to leave K whenever the construction leads to a node in K. There may be finitely many cycles in the part of the path that lies within K, but they can be pruned off to leave a simple directed cycle γ'', which satisfies the conditions (6.4.21) and contains at least one node which is not in K.

Since the cycle γ'' satisfies the conditions (6.4.21), it can be used in place of the cycle γ' in the proof of Theorem (6.4.18). By the argument in the first paragraph of the present proof, we conclude that $|x_m| = c_{j_r} =$ constant for all $P_m \in \Gamma_{\text{out}}(P_{j_r})$ for all $P_{j_r} \in \gamma''$. Therefore, every node in γ'' is in K, a contradiction to the conclusion that γ'' contains at least one node that is not in K. This shows that there can be no node of $\Gamma(A)$ that is not in K.

If γ is *any* nontrivial (simple directed) cycle in $\Gamma(A)$, it will automatically satisfy the conditions (6.4.21) because all its nodes are in K. It may therefore be used in place of γ' in the proof of Theorem (6.4.18), and hence it may be used in place of γ' in (6.4.28). This is the desired conclusion: The boundary of every set (6.4.27) passes through λ. □

6.4.29 Corollary. If $A \in M_n$, then either of the following conditions is sufficient for A to be invertible:

(a) A is weakly irreducible and

$$\prod_{P_i \in \gamma} |a_{ii}| > \prod_{P_i \in \gamma} R_i'$$

for every nontrivial cycle $\gamma \in C(A)$; or

(b) A is irreducible and

$$\prod_{P_i \in \gamma} |a_{ii}| \geq \prod_{P_i \in \gamma} R_i'$$

for every nontrivial cycle $\gamma \in C(A)$ with strict inequality for at least one cycle.

Problems

1. Show that if the matrix $A = [a_{ij}]$ satisfies Brauer's condition (6.4.11b) for invertibility, then $|a_{ii}| > R_i'$ for all but at most one value of $i = 1, \ldots, n$. Thus, Brauer's condition is only slightly weaker than the Levy–Desplanques condition (6.1.10a) of strict diagonal dominance. How is this related to (6.1.11)?

2. Show that $A = \begin{bmatrix} 2 & 3 \\ 1 & 3 \end{bmatrix}$ is invertible by both conditions (6.4.11) but neither the Levy–Desplanques condition (6.1.10a) nor (6.1.11) guarantee invertibility. What about the column form of (6.1.11)?

3. Show that every irreducible matrix $A \in M_n$ with $n \geq 2$ is weakly irreducible. Give an example of a weakly irreducible matrix that is not irreducible.

4. Provide the details for the proof of Corollary (6.4.29). *Hint:* Use the same arguments as in (6.1.10) and (6.2.6).

5. Show that $A \in M_n$ is weakly irreducible if and only if A is *not* permutation similar to a block triangular (0.9.4) matrix one of whose diagonal blocks is 1-by-1.

Further Reading. For more details about inclusion regions and many references to the original literature, see R. Brualdi, "Matrices, Eigenvalues, and Directed Graphs," *Lin. Multilin. Alg.* 11 (1982), 143–165.

CHAPTER 7

Positive definite matrices

7.0 Introduction

A class of Hermitian matrices with a special positivity property arises naturally in many applications. Hermitian (and, in particular, real symmetric) matrices with this positivity property also provide one generalization to matrices of the notion of a positive number. This observation often provides insight into the properties and applications of positive definite matrices. The following are examples of ways in which these special Hermitian matrices arise.

Hessians, minimization, and convexity

Let $f(x)$ be a smooth real-valued function on some domain $D \subset \mathbf{R}^n$. If $y = [y_i]$ is an interior point of D, then Taylor's theorem says that

$$f(x) = f(y) + \sum_{i=1}^{n} (x_i - y_i) \left.\frac{\partial f}{\partial x_i}\right|_y + \sum_{i,j=1}^{n} (x_i - y_i)(x_j - y_j) \left.\frac{\partial^2 f}{\partial x_i \, \partial x_j}\right|_y + \cdots$$

for points $x \in D$ which are near y. If y is a *critical point* of f, then all the first-order partial derivatives vanish at y and we have the expression

$$f(x) - f(y) = \sum_{i,j=1}^{n} (x_i - y_i)(x_j - y_j) \left.\frac{\partial^2 f}{\partial x_i \, \partial x_j}\right|_y + \cdots = (x - y)^T H(f; y)(x - y) + \cdots$$

for the behavior of f near y. The n-by-n matrix

$$H(f;y) \equiv \left[\frac{\partial^2 f}{\partial x_i \, \partial x_j} \bigg|_y \right]$$

is called the *Hessian* of f at y; it is a symmetric matrix because of equality of the mixed partial derivatives of f. If the quadratic form

$$z^T H(f;y) z, \qquad z \neq 0, \qquad z \in \mathbf{R}^n \tag{7.0.1}$$

is always positive, then y is a *relative minimum* for f. If this quadratic form is always negative, then y is a *relative maximum* for f. Of course, this quadratic form might not have a definite sign for all nonzero $z \in \mathbf{R}^n$, in which case the nature of the critical point y is not determined. In the case $n=1$, these criteria are just the usual second derivative test for a relative minimum or a maximum. The third possibility occurs for $n=1$ only at a point of inflection; when $n>1$, the situation can be much more complicated.

If the quadratic form (7.0.1) is nonnegative at all points of D (not just at the critical points of f), then f is a *convex function* in D. This is a direct generalization of the familiar situation when $n=1$.

Variance-covariance matrices

Let $X_1, X_2, \ldots, X_n$ be real or complex random variables with finite second moments on some probability space with expectation functional E, and suppose that $\mu_i = E(X_i)$ are the respective means. The *covariance matrix* of the random vector $X = (X_1, \ldots, X_n)^T$ is the matrix $A = [a_{ij}]$ in which

$$a_{ij} = E[(\bar{X}_i - \bar{\mu}_i)(X_j - \mu_j)], \qquad i, j = 1, \ldots, n$$

It is apparent that A is Hermitian, and one computes easily that if $z = [z_i] \in \mathbf{C}^n$, then

$$z^* A z = E\left[\sum_{i,j=1}^{n} \bar{z}_i (\bar{X}_i - \bar{\mu}_i) z_j (X_j - \mu_j) \right]$$

$$= E \left| \sum_{i=1}^{n} z_i (X_i - \mu_i) \right|^2 \geq 0$$

The only properties of the expectation functional that are involved in this observation are its linearity, homogeneity, and nonnegativity; that is, $E[Y] \geq 0$ whenever Y is a nonnegative random variable.

The same observation can be made without recourse to probabilistic language. If one has a family of complex valued functions $f_1, f_2, \ldots, f_n$ on the real line, if g is a real-valued function, and if all the integrals

$$a_{ij} = \int_{-\infty}^{\infty} \bar{f}_i(x) f_j(x) g(x) \, dx, \qquad i, j = 1, \ldots, n$$

are defined and converge, then the matrix $A = [a_{ij}]$ is obviously Hermitian. One computes easily that

$$z^* A z = \sum_{i,j=1}^{n} \int_{-\infty}^{\infty} \bar{z}_i \bar{f}_i(x) z_j f_j(x) g(x) \, dx$$

$$= \int_{-\infty}^{\infty} \left| \sum_{i=1}^{n} z_i f_i(x) \right|^2 g(x) \, dx$$

so this quadratic form will always be nonnegative if $g(x)$ is a nonnegative function.

Algebraic moments of nonnegative functions

Let $f(x)$ be an absolutely integrable real-valued function on the unit interval $[0, 1]$ and consider the numbers

$$a_k \equiv \int_0^1 x^k f(x) \, dx \tag{7.0.2}$$

The sequence $a_0, a_1, a_2, \ldots$ is said to be a *Hausdorff moment sequence*, and it is naturally associated with the real quadratic form

$$\sum_{j,k=0}^{n} a_{j+k} z_j z_k = \sum_{j,k=0}^{n} \int_0^1 x^{j+k} z_j z_k f(x) \, dx = \int_0^1 \left(\sum_{k=0}^{n} z_k x^k \right)^2 f(x) \, dx \tag{7.0.3}$$

If we set $A \equiv [a_{i+j}]$, then A will be a symmetric real matrix and we shall have $z^T A z \geq 0$ for all $z \in \mathbf{R}^{n+1}$ if $f(x) \geq 0$ for all $x \in [0, 1]$. This is true for each $n = 1, 2, \ldots$. A matrix with the structure of A (i.e., the elements a_{ij} are a function only of $i+j$) is called a *Hankel matrix*, whether or not its quadratic form is nonnegative. See Section (0.9.8).

Trigonometric moments of nonnegative functions

Let $f(\theta)$ be an absolutely integrable real-valued function on $[0, 2\pi]$ and consider the numbers

$$a_k \equiv \int_0^{2\pi} e^{ik\theta} f(\theta) \, d\theta, \qquad k = \pm 1, \pm 2, \ldots \tag{7.0.4}$$

The sequence $a_0, a_1, a_{-1}, a_2, a_{-2}, \ldots$ is said to be a *Toeplitz moment sequence*, and it is naturally associated with the quadratic form

$$\sum_{j,k=0}^{n} a_{j-k} z_j \bar{z}_k = \sum_{j,k=0}^{n} \int_0^{2\pi} e^{i(j-k)\theta} z_j \bar{z}_k f(\theta)\, d\theta$$
$$= \int_0^{2\pi} \left| \sum_{k=0}^{n} z_k e^{ik\theta} \right|^2 f(\theta)\, d\theta \qquad (7.0.5)$$

If we set $A \equiv [a_{i-j}]$, then A will be a Hermitian matrix and we shall have $z^*Az \geq 0$ for all $z \in \mathbf{C}^{n+1}$ if $f(\theta) \geq 0$ for all $\theta \in [0, 2\pi]$. This is true for each $n = 1, 2, \ldots$. A matrix which has the structure of A (i.e., the elements a_{ij} are a function only of $i-j$) is called a *Toeplitz matrix*, whether or not its quadratic form is nonnegative. See Section (0.9.7). It is a fact (Bochner's theorem) that nonnegativity of the quadratic form (7.0.5) is both necessary and sufficient for the numbers a_k to be generated by a slight modification of the formula (7.0.4) (in which a nonnegative measure $d\mu$ replaces $f(\theta)\, d\theta$).

Discretization and difference schemes for numerical solution of differential equations

Suppose we have a two-point boundary value problem of the form

$$-y''(x) + \sigma(x) y(x) = f(x), \qquad 0 \leq x \leq 1$$
$$y(0) = \alpha$$
$$y(1) = \beta$$

where α and β are given real constants, and $f(x)$ and $\sigma(x)$ are given real-valued functions. If we discretize this problem and look only for the values of $y(kh) \equiv y_k$, $k = 0, 1, \ldots, n+1$, and if we use a divided difference approximation to the derivative term

$$y''(x) \cong \frac{y((k+1)h) - 2y(kh) + y((k-1)h)}{h^2} = \frac{y_{k+1} - 2y_k + y_{k-1}}{h^2}$$

we obtain a system of linear equations

$$\frac{-y_{k+1} + 2y_k - y_{k-1}}{h^2} + \sigma_k y_k = f_k, \qquad k = 1, 2, \ldots, n$$
$$y_0 = \alpha$$
$$y_{n+1} = \beta$$

Here we have taken $h = 1/(n+1)$ for n a positive integer, $y_k = y(kh)$, $\sigma_k = \sigma(kh)$, and $f_k = f(kh)$. The boundary conditions can be incorporated into the first ($k=1$) and last ($k=n$) equations to give the system

$$(2 + h^2\sigma_1) y_1 - y_2 = h^2 f_1 + \alpha$$
$$-y_{k-1} + (2 + h^2\sigma_k) y_k - y_{k+1} = h^2 f_k, \qquad k = 2, 3, \ldots, n-1$$
$$-y_{n-1} + (2 + h^2\sigma_n) y_n = h^2 f_n + \beta$$

which can be written more compactly as $Ay = w$, where $y = [y_k] \in \mathbf{R}^n$, $w = [h^2 f_1 + \alpha, h^2 f_2, \ldots, h^2 f_{n-1}, h^2 f_n + \beta]^T \in \mathbf{R}^n$, and $A \in M_n$ is the tridiagonal matrix

$$A = \begin{bmatrix} 2+h^2\sigma_1 & -1 & & & 0 \\ -1 & 2+h^2\sigma_2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2+h^2\sigma_{n-1} & -1 \\ 0 & & & -1 & 2+h^2\sigma_n \end{bmatrix} \qquad (7.0.6)$$

Notice that A is a real symmetric tridiagonal matrix regardless of the values of $\sigma(x)$, but if we want to be able to solve $Ay = w$ for any given right-hand side, then we must impose some condition on $\sigma(x)$ to ensure that A is nonsingular.

It is easy to compute the real quadratic form associated with A:

$$x^T A x = \left[x_1^2 + \sum_{i=1}^{n-1} (x_i - x_{i+1})^2 + x_n^2 \right] + h^2 \sum_{i=1}^{n} \sigma_i x_i^2$$

The first group of three terms is nonnegative and can vanish only if the components of x are all equal, and equal to zero. If $\sigma(x) \geq 0$, then the last sum is nonnegative and

$$x^T A x \geq \left[x_1^2 + \sum_{i=1}^{n-1} (x_i - x_{i+1})^2 + x_n^2 \right] \geq 0 \qquad (7.0.7)$$

If A is singular, then there is some nonzero vector $\hat{x} \in \mathbf{R}^n$ such that $A\hat{x} = 0$, and hence $\hat{x}^T A \hat{x} = 0$. But then the central group of terms in (7.0.7) must vanish, which implies that $\hat{x} = 0$. Thus, if $\sigma(x) \geq 0$, the matrix A is nonsingular and the discretized boundary value problem can be solved for arbitrary boundary conditions α and β.

This is a typical situation in the study of numerical solutions of ordinary or partial differential equations. For computational stability it is desirable to design a discretization of a differential equation problem that leads to a system of linear equations $Ay = w$ in which A is positive definite, and it is usually possible to do so when the differential equations are elliptic.

Matrices with the special positivity property illustrated in these examples are the object of study in this chapter. These matrices arise in

many applications: in harmonic analysis, in complex analysis, in the theory of vibrations of mechanical systems, and in other areas of matrix theory such as the singular value decomposition and the solution of linear least-squares problems.

Problems

1. If the sequence a_k is generated by the formula (7.0.2) with a nonnegative function f, show that the quadratic forms

$$\sum_{i,j=1}^{n} a_{i+j+1} z_i z_j \quad \text{and} \quad \sum_{i,j=1}^{n} \{a_{i+j} - a_{i+j+1}\} z_i z_j, \qquad z = [z_i] \in \mathbf{R}^n$$

are both nonnegative.

2. Make a sketch illustrating which diagonals are constant in a Hankel matrix. Do the same for a Toeplitz matrix.

3. Show that the matrix A in (7.0.6) is always irreducible, and that it is irreducibly diagonally dominant if $\sigma(x) \geq 0$. Use Corollary (6.2.27) to show that A is nonsingular and that all the eigenvalues of A are positive.

Further Readings. For a short survey of facts about real positive definite matrices see C. R. Johnson, "Positive Definite Matrices," *Amer. Math. Monthly* 77(1970), 259–264. Other surveys that focus on different areas involving positive definite matrices and contain numerous references are O. Taussky, "Positive Definite Matrices," pp. 309–319 of *Inequalities*, ed. O. Shisha, Academic Press, New York, 1967; and O. Taussky, "Positive Definite Matrices and Their Role in the Study of the Characteristic Roots of General Matrices," *Advan. Math.* 2(1968), 175–186.

7.1 Definitions and properties

An n-by-n Hermitian matrix A is said to be *positive definite* if

$$x^*Ax > 0 \quad \text{for all nonzero } x \in \mathbf{C}^n \tag{7.1.1}$$

If the strict inequality required in (7.1.1) is weakened to $x^*Ax \geq 0$, then A is said to be *positive semidefinite*. Implicit in these defining inequalities is the observation that if A is Hermitian, the left-hand side of (7.1.1) is always a real number. Of course, if A is positive definite, then it is also positive semidefinite.

Exercise. What do positive definite and positive semidefinite mean when $n = 1$?

Exercise. Show that if $A \in M_n$ and if x^*Ax is real for all $x \in \mathbf{C}^n$, then A is Hermitian. Thus, the assumption that A is Hermitian is not necessary in the definition of positive definiteness. It is customary, however. *Hint:* Write $A = B + iC$ with B and C Hermitian.

Exercise. Show that if $A \in M_n$ is a real matrix and if x^TAx is positive for all nonzero $x \in \mathbf{R}^n$, then A need not be symmetric, and hence it need not be positive definite. *Hint:* Consider a real skew-symmetric matrix A and compute $(x^TAx)^T$. What is x^TAx in this case? What about x^*Ax for nonreal x?

Exercise. Show that $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ is positive semidefinite but not positive definite.

Exercise. Show that if $A = [a_{ij}] \in M_n$ is positive definite, then so are $\bar{A} = [\bar{a}_{ij}]$, A^T, A^*, and A^{-1}. *Hint:* If $Ay = x$, $x^*A^{-1}x = y^*A^*y$.

Similarly, the terms *negative definite* and *negative semidefinite* may be defined for A by reversing the inequalities in the definitions of positive definite and positive semidefinite or, equivalently, by saying that $-A$ is positive definite or positive semidefinite, respectively. Thus, any statement about negative definite matrices mirrors a statement about positive definite matrices. If a Hermitian matrix falls into none of the aforementioned classes [i.e., if the left-hand side of (7.1.1) takes on both positive and negative values], it is said to be *indefinite*.

Several immediate observations may be made about positive definite matrices, and each has an analog for positive semidefinite matrices.

7.1.2 Observation. Any principal submatrix of a positive definite matrix is positive definite.

Proof: Let S be a proper subset of $\{1, 2, \ldots, n\}$ and denote by $A(S)$ the matrix resulting from deleting the rows and columns complementary to those indicated by S from the positive definite matrix $A \in M_n$. Then $A(S)$ is a principal submatrix of A, and all principal submatrices arise in this way; recall that the number $\det A(S)$ is a principal minor of A. Let $x \in \mathbf{C}^n$ be a nonzero vector with arbitrary entries in the components indicated by S and zero entries elsewhere. Let $x(S)$ denote the vector obtained from x by deleting the (zero) components complementary to S, and observe that

$$x(S)^*A(S)x(S) = x^*Ax > 0$$

Since $x(S) \neq 0$ is arbitrary, this means that $A(S)$ is positive definite. □

Exercise. Show that the diagonal entries of a positive definite matrix are positive real numbers.

7.1.3 **Observation.** The sum of any two positive definite matrices of the same size is positive definite. More generally, any nonnegative linear combination of positive semidefinite matrices is positive semidefinite.

Proof: Let A and B be positive semidefinite, let $a, b \geq 0$, and observe that $x^*(aA+bB)x = a(x^*Ax)+b(x^*Bx) \geq 0$ for any $x \in \mathbf{C}^n$. The case of more than two summands is treated in the same way. If the coefficients are positive, if A and B are positive definite, and if the vector x is nonzero, then every term in the sum is positive, so a positive linear combination of positive definite matrices is positive definite. □

Thus, the set of positive definite matrices is a *positive cone* in the vector space of all matrices.

7.1.4 **Observation.** Each eigenvalue of a positive definite matrix is a positive real number.

Proof: Let A be positive definite, let $\lambda \in \sigma(A)$, let x be an eigenvector of A associated with λ, and calculate $x^*Ax = x^*\lambda x = \lambda x^*x$. Therefore, $\lambda = (x^*Ax)/x^*x$ is positive since it is a ratio of two positive numbers. □

7.1.5 **Corollary.** The trace, the determinant, and all principal minors of a positive definite matrix are positive.

Proof: The trace and determinant are just the sum and product of the eigenvalues. The rest follows from (7.1.2).

Exercise. Show that the eigenvalues, trace, determinant, and principal minors of a positive semidefinite matrix are all nonnegative.

Exercise. Show that the eigenvalues and trace of an n-by-n negative definite matrix are negative, but the determinant is negative for odd n and positive for even n.

Exercise. Show that if $A = [a_{ij}] \in M_2$ is positive definite, then $a_{11}a_{22} > |a_{12}|^2$. *Hint:* Use $\det A > 0$. Deduce that if $A \in M_n$ is positive definite, then

$$a_{ii}a_{jj} > |a_{ij}|^2$$

for all $i, j = 1, 2, \ldots, n$. Show that "$>$" must be replaced by "$\geq$" in this inequality if one assumes only that A is positive semidefinite.

7.1.6 **Observation.** Let $A \in M_n$ be positive definite. If $C \in M_{n,m}$, then C^*AC is positive semidefinite. Furthermore, $\operatorname{rank}(C^*AC) = \operatorname{rank}(C)$, so that C^*AC is positive definite if and only if C has rank m.

Proof: First note that C^*AC is Hermitian. For any $x \in \mathbf{C}^m$ we have $x^*C^*ACx = y^*Ay \geq 0$, where $y \equiv Cx$ and the inequality follows from the positive definiteness of A. Thus, C^*AC is positive semidefinite. Further, note that $x^*C^*ACx > 0$ if and only if $Cx \neq 0$ because A is positive definite. The statement about rank (and thus about the positive definiteness of C^*AC) would follow if we knew that $C^*ACx = 0$ if and only if $Cx = 0$ because this would mean that C^*AC and C have the same null space (and hence they also have the same rank). If $Cx = 0$, then obviously $C^*ACx = 0$. Conversely, if $C^*ACx = 0$, then $x^*C^*ACx = 0$ and (using the positive definiteness of A as before) we conclude that $Cx = 0$. □

Exercise. If $A \in M_n$ is positive semidefinite and not positive definite, and if $C \in M_n$, show that C^*AC is always positive semidefinite and not positive definite. If $C \in M_{n,m}$ with $n \neq m$, show by example that C^*AC may be positive definite even if $A \in M_n$ is singular.

Exercise. Show that the cone of positive (semi)definite matrices is invariant under *congruence. See (4.5.4).

Exercise. Let $A \in M_n$ be Hermitian. Show that A is positive (semi)definite if and only if there is a nonsingular matrix $C \in M_n$ such that C^*AC is positive (semi)definite.

What happens if one drops the requirement that A be Hermitian and uses only real vectors in the defining quadratic form (7.1.1)? If A is a matrix with real entries and if $x \in \mathbf{R}^n$, then x^TAx is real and we may still ask which matrices have $x^TAx > 0$ for all $x \neq 0$ (even if A is not symmetric). If A is a matrix with complex entries, or if $x \in \mathbf{C}^n$ is allowed, we might replace (7.1.1) with

$$\operatorname{Re}(x^*Ax) > 0 \quad \text{for all nonzero} \quad x \in \mathbf{C}^n \tag{7.1.1'}$$

Define the *Hermitian part* of A to be

$$H(A) \equiv \tfrac{1}{2}(A + A^*) \tag{7.1.7}$$

When $n = 1$ this is just the real part of the complex number A.

Exercise. Show that (7.1.1′) holds if and only if $H(A)$ is positive definite.

Exercise. Show that for any $A \in M_n$, $A = H(A) + S(A)$, where $S(A) = \frac{1}{2}(A - A^*)$ is the *skew-Hermitian part* of A.

Problems

1. Let $A \in M_n$ be positive semidefinite and $x \in \mathbf{C}^n$. Show that $x^*Ax = 0$ if and only if $Ax = 0$. Conclude that a positive semidefinite matrix $A \in M_n$ has rank n if and only if it is positive definite. *Hint:* Consider the quadratic polynomial $p(t) = (x+ty)^*A(x+ty)$, $t \in \mathbf{R}$. If $x^*Ax = 0$, show that $p(t) \geq 0$ for all t, $p(0) = 0$ and $dp/dt = 0$ at $t = 0$. Conclude that $y^*Ax = 0$ for all $y \in \mathbf{C}^n$ and hence that $Ax = 0$.

2. Show that if a positive semidefinite matrix has a zero entry on the main diagonal, then the entire row and column to which it belongs must be zero.

3. Show that if the main diagonal entries of a positive definite matrix are all $+1$, then all entries of the matrix are bounded by 1 in absolute value. Can equality occur?

4. Show that a positive semidefinite matrix A is of rank 1 if and only if A is of the form $A = xx^*$ for some nonzero vector $x \in \mathbf{C}^n$.

5. Let

$$A = [a_{ij}] \in M_n$$

be positive definite. Show that the matrix $[a_{ij}/(a_{ii}a_{jj})^{1/2}]$ is positive definite, that all its main diagonal entries are $+1$, and that all its entries are bounded by 1 in absolute value. Such a matrix is called a *correlation matrix*. *Hint:* Find a congruence by a certain real diagonal matrix.

6. If A has real entries, show that the requirement $x^TAx > 0$ for all nonzero $x \in \mathbf{R}^n$ depends only on $H(A)$.

7. Show that statements analogous to (7.1.2), (7.1.3), (7.1.4), and (7.1.6) hold for matrices $A \in M_n(\mathbf{C})$ such that $H(A)$ is positive definite.

8. A function $f: \mathbf{R} \to \mathbf{C}$ is said to be a *positive definite function* if the matrix $[f(x_i - x_j)] \in M_n$ is positive semidefinite for all choices of points $\{x_1, x_2, \ldots, x_n\} \subset \mathbf{R}$ and all $n = 1, 2, \ldots$. Show that $f(-x) = \bar{f}(x)$ for all $x \in \mathbf{R}$. Use the fact that the determinant of a positive semidefinite matrix is nonnegative to show that if f is a positive definite function, then

(a) $f(0) \geq 0$, $n = 1$;

(b) f is a bounded function, and $|f(x)| \leq f(0)$ for all $x \in \mathbf{R}$, $n = 2$;

(c) If f is continuous at 0, then it is continuous everywhere, $n = 3$.

9. If $f_1(x), f_2(x), \ldots, f_n(x)$ are positive definite functions and if $a_1, a_2, \ldots, a_n$ are nonnegative real numbers, show that the function $f(x) = a_1 f_1(x) + \cdots + a_n f_n(x)$ is a positive definite function.

10. Show that the function e^{itx} is a positive definite function for each given $t \in \mathbf{R}$. Use Problem 9 to show that $f(x) = a_1 e^{it_1 x} + \cdots + a_n e^{it_n x}$ is a positive definite function for any choice of points $t_1, \ldots, t_n \in \mathbf{R}$ and any nonnegative real numbers $a_1, \ldots, a_n$.

11. Prove that the function $\cos(x)$ is a positive definite function. *Hint:* $\cos(x) = (e^{ix} + e^{-ix})/2$.

12. Is $\sin(x)$ a positive definite function?

13. If $g(x)$ is a nonnegative and integrable function on $\mathbf{R}$, show that the function

$$f(x) = \int_{-\infty}^{\infty} e^{itx} g(t)\, dt$$

is a positive definite function. *Hint:* Use the definition.

14. Prove that the function $f(x) = 1/(1 - ix)$ is a positive definite function. *Hint:* Let $g(t) = e^{-t}$ for $t > 0$, $g(t) = 0$ for $t \leq 0$ in Problem 13.

15. It follows from Theorem (7.5.3) [see Problem 2 in Section (7.5)] that if $f(x)$ and $g(x)$ are positive definite functions, then so is $f(x)g(x)$. Show that if $f(x)$ is a positive definite function, then so are $\bar{f}(x)$ and $|f(x)|^2$, and use the latter fact to deduce from Problem 14 that the function $1/(1+x^2)$ is a positive definite function.

16. Use (7.0.2) and (7.0.3) with $f(x) \equiv 1$ to show that the matrix $A = [a_{ij}] \in M_n$ with $a_{ij} = 1/(i+j-1)$ for $i, j = 1, 2, \ldots, n$ is positive definite for all $n = 1, 2, \ldots$.

17. Show that the matrix $A = [a_{ij}] \in M_n$ with $a_{ij} = 1/(i+j)$ for $i, j = 1, 2, \ldots, n$ is positive definite for all $n = 1, 2, \ldots$. *Hint:* For all $x = [x_i] \in \mathbf{R}^n$,

$$\int_0^{\infty} \left(\sum_{k=1}^{n} x_k e^{-kt} \right)^2 dt \geq 0$$

Compute this integral.

18. Use (7.1.6) to show that the matrix $A = [a_{ij}] \in M_n$ with $a_{ij} = \min\{i, j\}$ is positive definite. *Hint:* What is this for $n = 4$? Consider the congruence C^*AC, where C is the real matrix

$$C = \begin{bmatrix} 1 & -1 & -1 & \cdots & -1 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \cdot & \ddots & \ddots & \vdots \\ & & & & 0 \\ 0 & 0 & \cdots & & 1 \end{bmatrix} \in M_n$$

Why is C nonsingular? Notice that the effect of this congruence is to subtract the first row (and column) of A from all the other rows (and

columns). Now observe the form of the lower right-hand $(n-1)$-by-$(n-1)$ submatrix of C^*AC and perform a suitable congruence on it to reduce it in the same way. Conclude that A is *congruent to I.

19. Use Problem 18 and a limiting argument to show that the kernel $K(s,t)=\min\{s,t\}$ is positive semidefinite on $[0,N]$ for any $N>0$, that is, that

$$\int_0^N\int_0^N K(s,t)\bar{f}(s)f(t)\,ds\,dt\geq 0 \tag{7.1.8}$$

for all continuous complex-valued functions $f(\bullet)$ on $[0,N]$. *Hint:* Express the integral as the limit of Riemann sums over partitions of $[0,N]$ with equally spaced points.

20. Prove the identity

$$\int_0^N\int_0^N \min\{s,t\}\bar{f}(s)f(t)\,ds\,dt=\int_0^N\left|\int_t^N f(s)\,ds\right|^2 dt$$

for all continuous complex valued functions $f(\bullet)$ on $[0,N]$ and use it to give an alternate proof of the assertion in Problem 19. This proof gives the stronger result that $K(s,t)=\min\{s,t\}$ is positive definite; that is, equality holds in (7.1.8) if and only if $f(t)=0$. *Hint:* Express the double integral as an iterated integral and integrate by parts.

7.2 Characterizations

There are several useful and simple characterizations of positive definite matrices.

7.2.1 **Theorem.** A Hermitian matrix $A\in M_n$ is positive semidefinite if and only if all of its eigenvalues are nonnegative. It is positive definite if and only if all of its eigenvalues are positive.

Proof: If each eigenvalue of A is positive, then for any nonzero $x\in\mathbf{C}^n$ we have

$$x^*Ax=x^*U^*DUx=y^*Dy=\sum_{i=1}^n d_i\bar{y}_iy_i=\sum_{i=1}^n d_i|y_i|^2>0$$

where $D=\operatorname{diag}(d_1,d_2,\ldots,d_n)$ is the diagonal matrix of eigenvalues of A, $y=Ux$, and U is unitary. The reverse implication is contained in Observation (7.1.4), and the positive semidefinite case is similar. □

Exercise. Show that a nonsingular $A\in M_n$ is positive definite if and only if A^{-1} is positive definite.

Exercise. Let $A\in M_n$ be positive semidefinite. Use (7.2.1) to show that A is positive definite if and only if rank $A=n$. Compare with Problem 1 of Section (7.1).

7.2.2 **Corollary.** If $A\in M_n$ is positive semidefinite, then so are all the powers A^k, $k=1,2,\ldots$.

Proof: If the eigenvalues of A are $\lambda_1,\ldots,\lambda_n$, then the eigenvalues of A^k are $\lambda_1^k,\ldots,\lambda_n^k$. □

7.2.3 **Corollary.** If $A=[a_{ij}]\in M_n$ is Hermitian and strictly diagonally dominant and if $a_{ii}>0$ for all $i=1,2,\ldots,n$, then A is positive definite.

Proof: This is part of Theorem (6.1.10). The conditions imply that each Geršgorin disc for A lies in the open right half-plane. Since the eigenvalues of a Hermitian matrix are all real, the eigenvalues of A must all be positive, and hence A is positive definite by the theorem. □

Exercise. If a Hermitian matrix A is *congruent to a strictly diagonally dominant matrix with positive diagonal entries, show that A is positive definite.

The next characterization is not very practical for computational determination of positive definiteness, but it can be of theoretical utility.

7.2.4 **Corollary.** Let $A\in M_n$ be Hermitian, and let

$$p_A(t)=t^n+a_{n-1}t^{n-1}+\cdots+a_{n-m}t^{n-m}$$

be the characteristic polynomial of A. Suppose that $0\leq m\leq n$ and $a_{n-m}\neq 0$. Then A is positive semidefinite if and only if $a_k\neq 0$ for all $n-m\leq k\leq n$ and $a_ka_{k+1}<0$ for $k=n-m,\ldots,n-1$. We define $a_n\equiv 1$.

Proof: The assertion is just that the leading coefficients a_k are nonzero and alternate strictly in sign. If this condition is met, $p_A(t)$ cannot have any negative zeroes; all the eigenvalues of A must therefore be nonnegative. Conversely, if A is positive semidefinite, denote its positive eigenvalues by $\lambda_1,\lambda_2,\ldots,\lambda_m$ (the remaining $n-m$ eigenvalues are all 0). By induction, one can show that the coefficients of the polynomials $(t-\lambda_1),(t-\lambda_1)(t-\lambda_2),\ldots,(t-\lambda_1)(t-\lambda_2)\cdots(t-\lambda_m)$ are all nonzero and alternate in sign. Multiplying by t^{n-m} gives $p_A(t)$. □

In order to facilitate the next characterization, we denote by A_i the leading principal submatrix of A determined by the first i rows and

columns, $A_i \equiv A(\{1,2,\ldots,i\})$, $i=2,\ldots,n$. We have already noted that if A is positive definite, then *all* principal minors of A are positive, and, in fact, the converse is valid when A is Hermitian. However, an even stronger statement may be made. Note that if A is Hermitian, so is each A_i, and therefore each A_i has a real determinant.

7.2.5 **Theorem.** If $A \in M_n$ is Hermitian, then A is positive definite if and only if $\det A_i > 0$ for $i=1,2,\ldots,n$. More generally, the positivity of *any* nested sequence of n principal minors of A (not just the leading principal minors) is necessary and sufficient for A to be positive definite.

Proof: Because of (7.1.5), we know that $\det A_i > 0$ for all $i=1,2,\ldots,n$ whenever A is positive definite. We use induction and the interlacing inequalities for a Hermitian matrix (4.3.8) to prove the converse. Since $\det A_1 > 0$ and A_1 is 1-by-1, A_1 is positive definite. If A_k is positive definite for some $k<n$, all the eigenvalues of A_k are positive and thus, by the interlacing inequalities, all the eigenvalues of A_{k+1} are positive except perhaps for the smallest eigenvalue. But the product of the eigenvalues of A_{k+1} is just $\det A_{k+1}$, which is assumed to be positive, so there cannot be just one negative eigenvalue for A_{k+1}. We conclude that even the smallest eigenvalue of A_{k+1} is positive, and hence A_{k+1} must be positive definite. Since $A_n = A$, we are done. For the case of a general nested sequence, just consider appropriate permutations of the rows and columns of A. □

Theorem (7.2.5) says that a Hermitian matrix is positive definite if (and only if) its leading principal minors are positive. Thus, noting (7.2.1), either of two sets of numbers associated with A may be checked in order to verify positive definiteness.

Exercise. Use (7.2.5) to show that the matrix

$$A = \begin{bmatrix} 5 & -1 & 3 \\ -1 & 2 & -2 \\ 3 & -2 & 3 \end{bmatrix}$$

is positive definite.

Exercise. Show that the leading principal minors of the symmetric matrix $\begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}$ are nonnegative, but it is not positive semidefinite.

Exercise. Let $A \in M_n$ be Hermitian and suppose that $\det A_1 > 0$, $\det A_2 > 0, \ldots, \det A_{n-1} > 0$, and $\det A_n \geq 0$. Show that A is positive semidefinite. *Hint:* What do the interlacing inequalities say about the eigenvalues of A_n as compared with those of A_{n-1}?

Exercise. Suppose that a Hermitian matrix $A \in M_n$ has positive diagonal entries and positive determinant. Consider the matrix

$$\begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 1 \\ 1 & 1 & t \end{bmatrix}$$

for suitable values of t to show that, by itself, this is not sufficient for positive definiteness. Show that if, in addition, some $(n-1)$-by-$(n-1)$ principal submatrix is diagonally dominant, then this condition *is* sufficient.

Exercise. Let $A \in M_n$ be Hermitian. Show that A is positive semidefinite if and only if there exist Hermitian matrices A_ϵ with $A_\epsilon \to A$ as $\epsilon \to 0$ such that every principal submatrix of A_ϵ has positive determinant. Conclude that if *all* principal minors of A are nonnegative, then A is positive semidefinite.

Every positive real number has a unique positive kth root for all $k=1,2,\ldots$. A similar result holds for positive definite matrices.

7.2.6 **Theorem.** Let $A \in M_n$ be positive semidefinite and let $k \geq 1$ be a given integer. Then there exists a unique positive semidefinite Hermitian matrix B such that $B^k = A$. We also have

(a) $BA = AB$ and there is a polynomial $p(t)$ such that $B = p(A)$;
(b) $\operatorname{rank} B = \operatorname{rank} A$, so B is positive definite if A is; and
(c) B is real if A is real.

Proof: We know that the Hermitian matrix A can be unitarily diagonalized as $A = U\Lambda U^*$ with $\Lambda = \operatorname{diag}(\lambda_1,\ldots,\lambda_n)$ and all $\lambda_i \geq 0$. We define $B = U\Lambda^{1/k}U^*$, where $\Lambda^{1/k} \equiv \operatorname{diag}(\lambda_1^{1/k},\ldots,\lambda_n^{1/k})$, and the unique nonnegative kth root is taken in each case. Clearly $B^k = A$ and B is Hermitian and positive semidefinite. Also, $AB = U\Lambda U^* U\Lambda^{1/k}U^* = U\Lambda\Lambda^{1/k}U^* = U\Lambda^{1/k}\Lambda U^* = U\Lambda^{1/k}U^*U\Lambda U^* = BA$, and B is positive semidefinite because all λ_i (and hence their kth roots) are nonnegative. The rank of B is just the number of nonzero λ_i terms, which is also the rank of A. If A is real and positive semidefinite, then we know that U may be chosen to be a real orthogonal matrix, so it is clear that B can be chosen to be real in this case. It remains only to consider the question of uniqueness.

Notice first that there is a polynomial $p(t)$ such that $p(A) = B$; we need only choose for $p(t)$ the Lagrange interpolating polynomial (0.9.11) for the set $\{(\lambda_1,\lambda_1^{1/k}),\ldots,(\lambda_n,\lambda_n^{1/k})\}$ to get $p(\Lambda) = \Lambda^{1/k}$ and $p(A) = p(U\Lambda U^*) = Up(\Lambda)U^* = U\Lambda^{1/k}U^* = B$. But then if C is any positive semidefinite Hermitian matrix such that $C^k = A$, we have $B = p(A) = p(C^k)$ so that

$CB = Cp(C^k) = p(C^k)C = BC$. Since B and C are commuting Hermitian matrices, they may be simultaneously unitarily diagonalized; that is, there is some unitary matrix V and diagonal matrices Λ_1 and Λ_2 with nonnegative diagonal entries such that $B = V\Lambda_1 V^*$ and $C = V\Lambda_2 V^*$. Then from the fact that $B^k = A = C^k$ we deduce that $\Lambda_1^k = \Lambda_2^k$. But since the nonnegative kth root of a nonnegative number is unique, we conclude that $(\Lambda_1^k)^{1/k} = \Lambda_1 = \Lambda_2 = (\Lambda_2^k)^{1/k}$ and $B = C$. □

The most useful case of the preceding theorem is for $k = 2$. The unique positive (semi)definite square root of the positive (semi)definite matrix A is usually denoted by $A^{1/2}$. Similarly, $A^{1/k}$ denotes the unique positive (semi)definite kth root of A for each $k = 1, 2, \ldots$.

Exercise. Determine $\begin{bmatrix} 5 & 3 \\ 3 & 2 \end{bmatrix}^{1/2}$.

Exercise. If A is positive definite, show that $(A^{1/2})^{-1} = (A^{-1})^{1/2}$.

7.2.7 Theorem. A matrix $B \in M_n$ is positive definite if and only if there is a nonsingular matrix $C \in M_n$ such that $B = C^*C$.

Proof: If B can be so written, then B is positive definite by (7.1.6). If B is positive definite, let $C = B^{1/2}$ to show that the asserted factorization can be achieved and that C can even be taken to be Hermitian. □

7.2.8 Corollary. A Hermitian matrix A is positive definite if and only if it is *congruent to the identity.

Proof: This is simply a restatement of (7.2.7).

Exercise. If $A \in M_n$ is positive definite, and if $A = C_1^*C_1$ and $A = C_2^*C_2$ with $C_1, C_2 \in M_n$, then show that $C_2 = VC_1$, where V is a unitary matrix. In particular, show that any solution C to $A = C^*C$ is of the form $C = VA^{1/2}$ with V unitary. *Hint:* Show that

$$A^{-1/2}C^*CA^{-1/2} = (CA^{-1/2})^*(CA^{-1/2}) = I$$

Sometimes it can be useful to know that the factorization $A = C^*C$ of a positive semidefinite matrix can be specialized somewhat. Every square matrix C has a QR factorization (2.6.1) and can be written as $C = QR$, where Q is unitary and R is an upper triangular matrix with the same rank as C. But then $A = C^*C = (QR)^*QR = R^*Q^*QR = R^*R$. If C is nonsingular, R may be chosen so that all its diagonal entries are positive (in fact, there is a unique factorization $C = QR$ of this type), and if C is real, both Q and R may be taken to be real. This establishes the following corollary, which gives the *Cholesky decomposition* of A.

7.2.9 Corollary. A matrix A is positive definite if and only if there exists a nonsingular lower triangular matrix $L \in M_n$ with positive diagonal entries such that $A = LL^*$. If A is real, L may be taken to be real.

Let $\{v_1, \ldots, v_k\}$ be a set of k given vectors in an inner product space V, and let $\langle \cdot, \cdot \rangle$ be a given inner product on V. The *Gram matrix* of the vectors $\{v_1, \ldots, v_k\}$ with respect to the inner product $\langle \cdot, \cdot \rangle$ is the matrix $G = [g_{ij}] \in M_k$ defined by $g_{ij} = \langle v_j, v_i \rangle$. Our final characterization of positive semidefinite matrices is that they are always Gram matrices (7.2.11).

7.2.10 Theorem. Let $G \in M_k$ be the Gram matrix of the vectors $\{w_1, \ldots, w_k\} \subset \mathbf{C}^n$ with respect to a given inner product $\langle \cdot, \cdot \rangle$, and let $W = [w_1\ w_2 \ldots w_k] \in M_{n,k}$. Then

(a) G is positive semidefinite;
(b) G is nonsingular if and only if the vectors $w_1, \ldots, w_k$ are independent;
(c) There exists a positive definite matrix $A \in M_n$ such that $G = W^*AW$; and
(d) rank G = rank W = maximum number of independent vectors in the set $\{w_1, \ldots, w_k\}$.

Proof: If $G = [g_{ij}]$ with $g_{ij} = \langle w_j, w_i \rangle$, then G is Hermitian because an inner product is Hermitian, and

$$x^*Gx = \sum_{i,j=1}^{k} g_{ij}\bar{x}_i x_j = \sum_{i,j=1}^{k} \langle w_j, w_i \rangle \bar{x}_i x_j = \sum_{i,j=1}^{k} \langle x_j w_j, x_i w_i \rangle$$

$$= \left\langle \sum_{j=1}^{k} x_j w_j, \sum_{i=1}^{k} x_i w_i \right\rangle = \left\| \sum_{i=1}^{k} x_i w_i \right\|^2 \geq 0$$

where $\|\cdot\|$ is the norm derived from the given inner product. By positive definiteness of the norm, equality can hold only if

$$\sum_{i=1}^{k} x_i w_i = 0$$

and this can happen for a nontrivial set of coefficients x_i only if the given vectors $\{w_i\}$ are dependent. If G is singular, there is some nonzero vector x such that $Gx = 0$ and hence $x^*Gx = 0$, which implies that the set $\{w_i\}$ is dependent. Conversely, if $x_1 w_1 + \cdots + x_k w_k = 0$ and $x = [x_i] \neq 0$, then we have shown that $x^*Gx = 0$, so G must be singular.

If $\{e_1, \ldots, e_n\}$ is the standard orthonormal basis of $\mathbf{C}^n$, then the matrix $A = (\langle e_j, e_i \rangle)$ is positive definite by (a) and (b). For any vectors $x, y \in \mathbf{C}^n$ we have

$$\langle y, x \rangle = \left\langle \sum_{j=1}^{n} y_j e_j, \sum_{i=1}^{n} x_i e_i \right\rangle = \sum_{i,j=1}^{n} \langle e_j, e_i \rangle \bar{x}_i y_j = x^* A y$$

so we have $g_{ij} = \langle w_j, w_i \rangle = w_i^* A w_j$ and hence $G = W^* A W$.

Finally, if $Gx = 0$, then $x^* G x = x^* W^* A W x = (Wx)^* A (Wx) = 0$, which implies that $Wx = 0$ since A is positive definite. Conversely, $Wx = 0$ implies that $Gx = W^* A (Wx) = 0$, so G and W have the same null space and hence they have the same rank. The column rank of W is the maximum number of independent vectors in the set $\{w_1, \ldots, w_k\}$. □

Exercise. The most common application of the theorem is to the case in which the given inner product is just the usual Euclidean inner product $\langle x, y \rangle = y^* x$. Show that $A = I$ in this case and deduce that the maximum number of independent vectors in a given set $\{w_1, \ldots, w_k\} \subset \mathbf{C}^n$ is exactly the rank of the matrix $G = [w_i^* w_j] \in M_k$.

7.2.11 **Corollary.** Let $A \in M_n$ be a given matrix. Then A is positive semidefinite with rank $r \le n$ if and only if there is a set of vectors $S = \{w_1, \ldots, w_n\} \subset \mathbf{C}^n$ containing exactly r independent vectors such that A is the Gram matrix of S with respect to the Euclidean inner product.

Proof: The "if" part has been treated in the theorem. For the "only if" part, use (7.2.6) to write $A = B^2$ with B positive semidefinite. The rank of B is the same as the rank of A and $A = B^2 = B^* B$ is the Gram matrix of the columns of B in the Euclidean inner product. □

Problems

1. Show that if A is a Hermitian matrix, then A^{2k} is positive semidefinite for all $k = 1, 2, \ldots$ and e^A is positive definite. See the exercises following (5.6.15).

2. If A is positive semidefinite, and if $p(t)$ is any polynomial such that $p(t) > 0$ for all $t \ge 0$, show that $p(A)$ is positive semidefinite. *Hint:* What are the eigenvalues of $p(A)$? How does this generalize Problem 1?

3. Use (7.2.5) to show that the matrix $A = [a_{ij}] \in M_n$ defined by $a_{ij} \equiv \min\{i, j\}$ is positive definite. *Hint:* Calculate $\det A_i$; subtract the first row from all the other rows, then do the same for the first column. What about $a_{ij} \equiv \max\{i, j\}$?

4. If A and B are positive definite, show that $\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$ is positive definite.

5. Give an example of a real square (non-Hermitian) matrix whose leading principal minors are positive but such that some eigenvalue has negative real part.

6. Provide the details for the general inequalities in (7.2.5). That is, show that the positivity of any nested sequence of n principal minors (nested by inclusion, not necessarily the leading principal minors) is sufficient for the positive definiteness of an n-by-n Hermitian matrix.

7. What are necessary and sufficient conditions for A to be negative definite (semidefinite) in terms of the signs of the minors?

8. Are there "square roots" of the positive semidefinite matrix A other than $A^{1/2}$? How many? Are there kth roots other than $A^{1/k}$? Are there non-Hermitian square roots? *Hint:* Consider $\begin{bmatrix} -1 & 1 \\ 0 & 1 \end{bmatrix}^2$.

9. If $B \in M_n$ is positive semidefinite and has rank m, show that there exists an m-by-n matrix C with rank m such that $B = C^* C$. In particular, note that a rank 1 positive semidefinite matrix may always be written in the form xx^* for some $x \in \mathbf{C}^n$.

10. Suppose $A \in M_n$ is positive semidefinite and has rank $r < n$. Show that A has an r-by-r positive definite principal submatrix.

11. Let $A \in M_n$ be Hermitian. Show that A is positive definite if and only if the classical adjoint adj A is positive definite and $\det A > 0$. If A is positive semidefinite, show that adj A is positive semidefinite and $\det A \ge 0$. *Hint:* Consider $A_\epsilon \equiv A + \epsilon I$, $\epsilon > 0$. Consider $A \equiv \operatorname{diag}(0, 0, -1)$ to show that one can have adj A positive semidefinite and $\det A \ge 0$ without having A positive semidefinite.

12. Let $r \in (0, 1)$ be given, and consider the real symmetric Toeplitz matrix $A = [a_{ij}] \in M_n$ defined by $a_{ij} = r^{|i-j|}$. Show that A is positive definite as follows: (a) If A_{ij} is the i, j minor of A, show that $\det A_{ij} = 0$ whenever $|i - j| \ge 2$. *Hint:* If $i = 1$ and $j > 2$, observe that the first column of A_{1j} is a multiple of the second column. (b) Let $D_n = \det A$. Show that $D_2 = 1 - r^2$ and use (a) to show that $D_{n+1} = D_n - r^2 D_n = (1 - r^2) D_n = (1 - r^2)^n$ by expanding according to cofactors of the first row. (c) Use (7.2.5) to conclude that A is positive definite.

13. Show that the matrix A in Problem 12 has an inverse that is real, symmetric, and tridiagonal, and that $(1 - r^2) A^{-1}$ has the entry $-r$ in every position of the superdiagonal and subdiagonal and has main diagonal entries $1, 1 + r^2, \ldots, 1 + r^2, 1$. *Hint:* Use Problem 12(a) to show that

A^{-1} is tridiagonal. Why must A^{-1} be symmetric? Now determine the elements of A^{-1} using $AA^{-1}=A^{-1}A=I$.

14. Let $\langle\cdot,\cdot\rangle$ be a given inner product on $\mathbf{C}^n$, let $\mathcal{B}=\{e_1,\ldots,e_n\}$ be the usual (with respect to the usual Euclidean inner product) orthonormal basis for $\mathbf{C}^n$, and let $G\in M_n$ denote the Gram matrix of $\mathcal{B}$ with respect to the given inner product $\langle\cdot,\cdot\rangle$. Show that

$$\langle x,y\rangle=y^*Gx \tag{7.2.12}$$

for all $x,y\in\mathbf{C}^n$. Conclude that a function $\langle\cdot,\cdot\rangle:\mathbf{C}^n\times\mathbf{C}^n\to\mathbf{C}$ is an inner product if and only if there is a positive definite matrix G such that (7.2.12) holds.

15. Recall the notion of a *dual norm* defined in (5.4.12). Let $\langle\cdot,\cdot\rangle$ be a given inner product on $\mathbf{C}^n$ and let $\|\cdot\|$ be a given norm on $\mathbf{C}^n$. The given norm is not necessarily induced by the given inner product. One can define the *dual norm of* $\|\cdot\|$ *with respect to the inner product* $\langle\cdot,\cdot\rangle$ as

$$\|x\|^D_{\langle\cdot,\cdot\rangle}\equiv\max_{\|y\|=1}|\langle x,y\rangle|$$

Notice that this is the usual dual of $\|\cdot\|$ if $\langle\cdot,\cdot\rangle$ is the Euclidean inner product. Does this extension of the notion of a dual norm produce any vector norms that we have not already generated by other means? *Hint:* Use Problem 14 to write $\langle x,y\rangle=y^*Gx$, and show that

$$\|x\|^D_{\langle\cdot,\cdot\rangle}=\|G^{-1}x\|^D\equiv(\|x\|_{G^{-1}})^D$$

16. Let $A\in M_n$ be given. Show that $\rho(A)<1$ if and only if there exists a positive definite matrix $B\in M_n$ such that $B-A^*BA$ is positive definite. *Hint:* If B is positive definite, let $C=B^{1/2}$. If

$$B-A^*BA=C^*C-(CA)^*(CA)$$

is positive definite, then for any nonzero $x\in\mathbf{C}^n$ we have

$$x^*[C^*C-(CA)^*(CA)]x>0$$

or $\|Cx\|_2>\|CAx\|_2$. Let $y=Cx$ to show that $\|y\|_2>\|CAC^{-1}y\|_2$ for all nonzero $y\in\mathbf{C}^n$ and conclude that $\|CAC^{-1}\|_2<1$. Thus, $\rho(A)=\rho(CAC^{-1})\leq\|CAC^{-1}\|_2<1$. Conversely, if $\rho(A)<1$ there exists a nonsingular $C\in M_n$ such that $\|CAC^{-1}\|_2<1$ [see Section (5.6), Problem 25] and the above argument can be reversed with $B\equiv C^*C$.

17. Let $A,B\in M_n$ be positive semidefinite and not both singular. Show that $\|A-B\|_2\leq\|A^2-B^2\|_2/[\lambda_{\min}(A)+\lambda_{\min}(B)]$. *Hint:* Let $E=A-B$ and let $x\in\mathbf{C}^n$ be a unit vector such that $Ex=\lambda x$ and $|\lambda|=\rho(E)=\|E\|_2$. Then $A^2-B^2=AE+EA-E^2$ and $\|A^2-B^2\|_2\geq|x^*(AE+EA-E^2)x|=|\lambda|(x^*Ax+x^*Bx)\geq|\lambda|(\lambda_{\min}(A)+\lambda_{\min}(B))$.

18. Let $A,B\in M_n$ be positive semidefinite and suppose A is positive definite. Use Problem 17 to show that

$$\|A^{1/2}-B^{1/2}\|_2\leq\|A^{-1/2}\|_2\|A-B\|_2 \tag{7.2.13}$$

and explain why this inequality implies that the function $f:C\to C^{1/2}$, defined on the set of positive semidefinite matrices in M_n, is continuous on the interior of this set, which is the open set of positive definite matrices. State and prove directly the inequality for the ordinary scalar square root function $f:t\to\sqrt{t}$ on $[0,\infty)$ that results from setting $n=1$ in (7.2.13).

7.3 The polar form and the singular value decomposition

We next develop two important related factorizations of complex matrices (not necessarily square) which depend heavily on positive definiteness.

7.3.1 Lemma. Let $A\in M_{m,n}$ with $m\leq n$ and rank $A=k\leq m$. There exists a unitary matrix $X\in M_m$, a diagonal matrix $\Lambda\in M_m$ with nonnegative diagonal entries $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_k>\lambda_{k+1}=\cdots=\lambda_m=0$, and a matrix $Y\in M_{m,n}$ with orthonormal rows such that $A=X\Lambda Y$. The matrix $\Lambda=\operatorname{diag}(\lambda_1,\ldots,\lambda_m)$ is always uniquely determined and $\{\lambda_1^2,\ldots,\lambda_m^2\}$ are the eigenvalues of AA^*. The columns of the matrix X are eigenvectors of AA^*. If AA^* has distinct eigenvalues, then X is determined up to a right diagonal factor $D=\operatorname{diag}(e^{i\theta_1},\ldots,e^{i\theta_n})$ with all $\theta_i\in\mathbf{R}$; that is, if $A=X_1\Lambda Y_1=X_2\Lambda Y_2$, then $X_2=X_1D$. Given X, the matrix Y is uniquely determined if rank $A=m$. If A is real, then X and Y may be taken to be real.

Proof: If $A=X\Lambda Y$ is a factorization of the asserted form, then $AA^*=X\Lambda YY^*\Lambda X^*=X\Lambda I\Lambda X^*=X\Lambda^2X^*$, so $X\Lambda^2X^*$ is a unitary diagonalization of the Hermitian matrix AA^*. If $X=[x_1\,x_2\ldots x_m]$ and if $\Lambda^2=\operatorname{diag}(\lambda_1^2,\ldots,\lambda_m^2)$, then $AA^*x_j=\lambda_j^2x_j$, $j=1,2,\ldots,m$, and the vectors $\{x_j\}$ are orthonormal. Because the diagonal entries of Λ are to be nonnegative and are to be arranged in nonincreasing order, Λ is uniquely determined by AA^*. If the numbers $\{\lambda_i^2\}$ are distinct, the corresponding normalized eigenvectors of AA^* are each determined up to a complex scalar factor of modulus 1, so if X_1 and X_2 are unitary matrices whose columns are eigenvectors of AA^*, we must have $X_2=X_1D$ with $D=\operatorname{diag}(d_1,\ldots,d_n)$ and all $|d_i|=1$.

Eigenvectors of AA^* corresponding to a multiple eigenvalue are not uniquely determined, however, but once they are chosen and orthonormalized so that the unitary matrix X is fixed, then $Y=\Lambda^{-1}X^*A$ is

uniquely determined if Λ is nonsingular, which is the case if $k = \text{rank } A = m$. One checks easily that $YY^* = \Lambda^{-1}X^*(AA^*X)\Lambda^{-1} = \Lambda^{-1}X^*X\Lambda^2\Lambda^{-1} = \lambda^{-1}\Lambda^2\Lambda^{-1} = I$, so this matrix Y has orthonormal rows.

It remains only to handle the case in which rank $A = k < m$. Since we want $Y = \Lambda^{-1}X^*A = \Lambda^{-1}(A^*X)^*$ when all $\lambda_i \neq 0$, we are led to define the jth row of Y to be the row vector y_j^*, where $y_j \equiv \lambda_j^{-1}(A^*x_j)$, $j = 1, \ldots, k$. Then

$$[\lambda_j^{-1}(A^*x_j)]^*[\lambda_k^{-1}(A^*x_k)] = x_j^*AA^*x_k/\lambda_j\lambda_k = x_j^*\lambda_k^2x_k/\lambda_j\lambda_k = x_j^*x_k\lambda_k/\lambda_j$$

which is 0 if $j \neq k$ and is 1 if $j = k$ since the vectors $\{x_j\}$ are orthonormal. The vectors $\{y_1, \ldots, y_k\}$ are an orthonormal set in $\mathbf{C}^n$, and $n \geq m > k$, so there exist $m-k$ additional (but not uniquely determined) orthonormal vectors $y_{k+1}, \ldots, y_m$ such that the matrix $Y^* \equiv [y_1\, y_2 \ldots y_k\, y_{k+1} \ldots y_m] \in M_{n,m}$ has m orthonormal columns.

Now notice that $X^*A = \Lambda Y$. The first k rows of both sides of this identity are equal by construction of the vectors y_j. The last $m-k$ rows are all 0 on the right because the last $m-k$ diagonal entries of Λ are 0; the last $m-k$ rows are all 0 on the left because if $AA^*x_j = 0$, then $0 = x_j^*AA^*x_j = (A^*x_j)^*(A^*x_j) = 0$ and hence $A^*x_j = 0$.

Finally, if A is real, then AA^* is real and has real eigenvalues, and hence the eigenvectors X may be taken to be real. The first k rows of Y, which are determined by X, are real by construction, and the $m-k$ orthonormal vectors that are added may be taken to be real. Thus, all the factors may be taken to be real if A is real. □

Every nonzero complex number z has a unique "polar representation" $z = pu$, where p is a positive real number and u is a complex number of modulus 1. Indeed, $p = |z|$ and $u = p^{-1}z = z/|z|$ if $z \neq 0$. If $z = 0$, then z can still be written in polar form with $p = 0$, but u is no longer uniquely determined. Indeed, u can be any complex number of modulus 1.

How does this generalize to a complex matrix $A \in M_n$? One answer is that $A = PU$ where P is positive (semi)definite and U is unitary. We can even generalize to the case in which A is not a square matrix.

7.3.2 **Theorem.** Let $A \in M_{m,n}$ with $m \leq n$. Then A may be written as

$$A = PU$$

where $P \in M_m$ is positive semidefinite, rank $P = \text{rank } A$, and $U \in M_{m,n}$ has orthonormal rows (that is, $UU^* = I$). The matrix P is always uniquely determined as $P = (AA^*)^{1/2}$, and U is uniquely determined when A has rank m. If A is real, then both P and U may be taken to be real.

Proof: Use (7.3.1) to write $A = X\Lambda Y = X\Lambda X^*XY$ and set $P = X\Lambda X^*$ and $U = XY$. Then P is positive semidefinite, and $UU^* = XYY^*X^* = XIX^* = XX^* = I$, so U has orthonormal rows. By the construction in (7.3.1), $P = (AA^*)^{1/2}$, and in general if $A = PU$, then $AA^* = PUU^*P = P^2$, so P must always be the (unique) positive semidefinite square root of AA^*. If A has rank m, then P is nonsingular, and $U = P^{-1}A$ is uniquely determined. As we saw in (7.3.1), however, if rank $A < m$, then the rows of Y corresponding to the 0 eigenvalues of P are not uniquely determined, so $U = XY$ need not be uniquely determined when rank $A < m$. □

An important special case follows immediately.

7.3.3 **Corollary.** If $A \in M_n$, then it may be written in the form

$$A = PU$$

where P is positive semidefinite and U is unitary. The matrix P is always uniquely determined as $P \equiv (AA^*)^{1/2}$; if A is nonsingular, then U is uniquely determined as $U \equiv P^{-1}A$. If A is real, then P and U may be taken to be real.

Exercise. Show that Theorem (7.3.2) may be proved using the following limit argument. If A is nonsingular, then set $P \equiv (AA^*)^{1/2}$, define $U \equiv P^{-1}A$, and check that $UU^* = I$. Thus, both P and U are uniquely determined. If A is singular, consider $A_\epsilon \equiv A + \epsilon I$, $\epsilon > 0$, and form $A_\epsilon = P_\epsilon U_\epsilon$ where both factors are uniquely determined. Use the selection principle (2.1.8) to obtain a sequence $\epsilon_k \to 0$ as $k \to \infty$ such that U_{ϵ_k} is entry-wise convergent to a unitary matrix U as $k \to \infty$. Since $P_{\epsilon_k} = A_{\epsilon_k}U_{\epsilon_k}^*$, we also have $P_{\epsilon_k} \to P$ and $A = PU$. Notice that this argument, while conceptually more economical than that given for (7.3.2) above, does not give a constructive procedure for obtaining the factors P and U when A is singular.

The factorization (7.3.2) is known as the *polar form* or *polar decomposition* of the matrix A. We note that both factors are unique if A has full rank.

Exercise. If $A \in M_{m,n}$ and $m \geq n$, show that it may be written as

$$A = WQ$$

where $W \in M_{m,n}$ has orthonormal columns (that is, $W^*W = I$) and $Q \in M_n$ is positive semidefinite. *Hint:* Factorize A^* by (7.3.2).

Exercise. Let $x \in \mathbf{C}^n$ be a given nonzero vector and let $A \equiv x \in M_{n,1}$. Show that the polar decomposition of A is $A = x = \|x\|_2 u$, where $u \equiv x/\|x\|_2$.

Thus, the polar decomposition may be thought of as a generalization to matrices of the convenient factorization $x=\|x\|_2(x/\|x\|_2)$ of nonzero vectors.

Exercise. Show that a square matrix A may be written both as $A=PU$ and as $A=WQ$, where $P=(AA^*)^{1/2}$ and $Q=(A^*A)^{1/2}$. These are sometimes called "left" and "right" polar decompositions of A. Show that the uniquely determined positive semidefinite factors P and Q are equal if and only if A is normal. It is a fact that if A is nonsingular, then the uniquely determined unitary factors U and W are always equal [exercise preceding Theorem (7.3.6)].

Exercise. Not every square matrix is normal; that is, it is not always true that $AA^*=A^*A$. But AA^* is always unitarily similar to A^*A. Use the polar decomposition (7.3.3) to prove this.

7.3.4 **Theorem.** Let $A\in M_n$, and let $A=PU$ be a polar decomposition. Then A is normal if and only if $PU=UP$.

Proof: If P and U commute, then $AA^*=PUU^*P^*=PP=P^2$ and $A^*A=U^*P^*PU=U^*P^2U=U^*UP^2=P^2$, and so A is normal. If A is normal, then $P^2=U^*P^2U$. Observe that P^2 and U^*P^2U are both positive semidefinite square matrices with obvious respective positive semidefinite square roots P and U^*PU. But Theorem (7.2.6) says that such a square root is unique, so $P=U^*PU$, or $UP=PU$. □

Our next goal is to deduce the *singular value decomposition* of an arbitrary (not necessarily square) matrix from (7.3.1).

7.3.5 **Theorem.** If $A\in M_{m,n}$ has rank k, then it may be written in the form

$$A=V\Sigma W^*$$

where $V\in M_m$ and $W\in M_n$ are unitary. The matrix $\Sigma=[\sigma_{ij}]\in M_{m,n}$ has $\sigma_{ij}=0$ for all $i\neq j$, and $\sigma_{11}\geq\sigma_{22}\geq\cdots\geq\sigma_{kk}>\sigma_{k+1,k+1}=\cdots=\sigma_{qq}=0$, where $q=\min\{m,n\}$. The numbers $\{\sigma_{ii}\}\equiv\{\sigma_i\}$ are the nonnegative square roots of the eigenvalues of AA^*, and hence are uniquely determined. The columns of V are eigenvectors of AA^* and the columns of W are eigenvectors of A^*A (arranged in the same order as the corresponding eigenvalues σ_i^2). If $m\leq n$ and if AA^* has distinct eigenvalues, then V is determined up to a right diagonal factor $D=\text{diag}(e^{i\theta_1},\ldots,e^{i\theta_n})$ with all $\theta_i\in\mathbf{R}$; that is, if $A=V_1\Sigma W_1^*=V_2\Sigma W_2^*$, then $V_2=V_1D$. If $m<n$, then W is never uniquely determined; if $n=m=k$ and V is given, then W is uniquely determined. If $n\leq m$, the uniqueness of V and W is determined by considering A^*. If A is real, then V, Σ, and W may all be taken to be real.

Proof: We assume without loss of generality that $m\leq n$ (otherwise, replace A by A^*). Use (7.3.1) to write $A=X\Lambda Y$ with $X,\Lambda\in M_m$ and $Y\in M_{m,n}$. Set $V\equiv X$, take $\Sigma\equiv[\Lambda\,|\,0]\in M_{m,n}$, and define $W\equiv[Y^*\,|\,S^*]\in M_n$ by requiring that the columns of W be an orthonormal set in $\mathbf{C}^n$. The columns of Y^* are already orthonormal, so if $m<n$, the columns of $S^*\in M_{n,(n-m)}$ may be chosen (but not uniquely) to make W be unitary. It is immediate that $V\Sigma W^*=X\Lambda Y=A$. The statements about uniqueness follow from the corresponding assertions in (7.3.1). □

The "diagonal entries" $\sigma_i=\sigma_{ii}$, $i=1,\ldots,q=\min\{m,n\}$ of Σ are known as the *singular values* of $A\in M_{m,n}$ (sometimes only the nonzero ones are so termed), and the columns of V and the columns of W are the (respectively, *left* and *right*) *singular vectors* of A. The factorization (7.3.5) is known as the *singular value decomposition* of A. The polar matrix P is the unique positive semidefinite square root of AA^*, and the singular values σ_i are the nonnegative square roots of the eigenvalues of AA^*, so the singular values of A are the same as the eigenvalues of the polar matrix P. While it is convenient to arrange the singular values in decreasing order, this is not a universal convention in the singular value decomposition; it is the *set* of singular values that is uniquely determined by A.

Notice that the singular value decomposition is a natural generalization to arbitrary matrices of the unitary diagonalization of normal matrices. For this reason, it is often the case that facts about eigenvalues of normal matrices generalize to statements about singular values of general matrices.

Exercise. Let $x\in\mathbf{C}^n$ be a given nonzero vector and let $A\equiv x\in M_{n,1}$. Show that a singular value decomposition of A is $A=x=V\Sigma W^*$, where $W=[1]\in M_1$, $\Sigma=[\|x\|_2,0,\ldots,0]^T\in M_{n,1}$, and $V=[v_1\ldots v_n]\in M_n$ has $v_1=x/\|x\|_2$, and $v_2,\ldots,v_n$ are $n-1$ arbitrary orthonormal vectors that are orthogonal to x.

If $A\in M_n$, the three factors V, Σ, and W in the singular value decomposition are all n-by-n matrices. If $A=PU$ is a polar decomposition of A, and if $P=V\Lambda V^*$ is a unitary diagonalization of P in which the (necessarily nonnegative) eigenvalues of P are arranged in nonincreasing order, then $A=PU=V\Lambda V^*U=(V)(\Lambda)(V^*U)=V\Lambda W^*$ is a singular value decomposition of A with $V=V$, $\Sigma=\Lambda$, and $W=U^*V$. Notice that $AA^*=V\Sigma W^*W\Sigma V^*=V\Sigma^2V^*$, so that the columns of V are eigenvectors of the Hermitian matrix AA^* with corresponding eigenvalues $\sigma_1^2,\ldots,\sigma_n^2$.

Similarly, $A^*A = W\Sigma V^* V\Sigma W^* = W\Sigma^2 W^*$, so the columns of W are eigenvectors of A^*A.

Exercise. If $A \in M_n$ is nonsingular, show that the following procedure yields a singular value decomposition $A = V\Sigma W^*$:

(a) Form the positive definite Hermitian matrix AA^* and compute a unitary diagonalization $AA^* = U\Lambda U^*$ by finding the (positive) eigenvalues $\{\lambda_i\}$ of AA^* and a corresponding set $\{u_i\}$ of normalized eigenvectors.
(b) Set $\Sigma = \Lambda^{1/2}$ and $V = U = [u_1 \dots u_n]$.
(c) Set $W \equiv A^* V \Sigma^{-1}$.

Show that W is unitary and $A = V\Sigma W^*$. *Hint:* Compute W^*W.

Exercise. If $A \in M_n$ is given (not necessarily nonsingular), show that the following procedure yields a singular value decomposition $A = V\Sigma W^*$:

(a) There exists some $c = c(A) > 0$ such that $A_\epsilon = A + \epsilon I$ is nonsingular for all positive $\epsilon < c$. Let $0 < \epsilon < c$.
(b) Use the procedure in the previous exercise to form a singular value decomposition $A_\epsilon = V_\epsilon \Sigma_\epsilon W_\epsilon^*$.
(c) Use the selection principle (2.1.8) and let $\epsilon \to 0$ through a sequence of values ϵ_k such that

$$\lim_{\epsilon_k \to 0} V_{\epsilon_k} = V \quad \text{and} \quad \lim_{\epsilon_k \to 0} W_{\epsilon_k} = W$$

both exist.
(d) Show that $A = V\Sigma W^*$, in which $\Sigma = \lim_{\epsilon \to 0} \Sigma_\epsilon$.

This argument, which can be used to prove the general singular value decomposition (7.3.5), guarantees that a singular value decomposition exists in general but does not give a constructive procedure for computing the factors in the singular value decomposition when A does not have full rank.

Exercise. Suppose $A \in M_n$ is nonsingular and that $A = PU$, $A = WQ$ are the left and right polar decompositions of A with positive definite $P, Q \in M_n$ and unitary $U, W \in M_n$. Show that $U = W$ always, but $P = Q$ if and only if A is normal. If A is singular, show that there exist left and right polar decompositions of A for which $U \neq W$. *Hint:* If $A = V\Sigma W^*$ is a singular value decomposition of A, then neither V nor W is uniquely determined but $A = (VW^*)(W\Sigma W^*) = (V\Sigma V^*)(VW^*)$; use the uniqueness part of (7.3.3). Consider $A = 0$ to show that the unitary factors in the two polar decompositions of A need not be the same if A is singular.

If $A \in M_n$ is normal, and if $A = V\Sigma W^*$ is a singular value decomposition, then $AA^* = A^*A$, and so AA^* and A^*A have the same eigenvectors. It does not follow from this that $V = W$ in a singular value decomposition of A, however, for then $A = V\Sigma V^*$ would necessarily be Hermitian (even positive semidefinite). If $A = U\Lambda U^*$ is a unitary diagonalization of A and if $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$, then each $\lambda_k = |\lambda_k| e^{i\theta_k}$ for some $\theta_k \in \mathbf{R}$; if $\lambda_k = 0$, choose $\theta_k \equiv 0$. If we set $D \equiv \text{diag}(e^{i\theta_1}, \dots, e^{i\theta_n})$ and $|\Lambda| \equiv \text{diag}(|\lambda_1|, \dots, |\lambda_n|)$, then $\Lambda = |\Lambda| D$ and $A = U\Lambda U^* = U|\Lambda| DU^* = (U)(|\Lambda|)(U\bar{D})^* \equiv V\Sigma W^*$ is a singular value decomposition of A with $V = U$, $\Sigma = |\Lambda|$, and $W = U\bar{D}$.

Thus, the singular values of a normal matrix are just the absolute values of the eigenvalues, the columns of V are eigenvectors of A, and the columns of W may be taken to be the same as the columns of V except that each is multiplied by a complex scalar of absolute value 1, which is determined by the corresponding eigenvalue. If A is Hermitian, then all the eigenvalues are real, $D = \bar{D}$, and $D = \text{diag}(\text{sgn}(\lambda_1), \dots, \text{sgn}(\lambda_n))$, where we set $\text{sgn}(0) \equiv 1$. If A is Hermitian and positive semidefinite, then $D = I$, $V = W = U$, and $\Lambda = \Sigma$.

One useful application of the Schur triangularization theorem (2.3.1) was to show that every square complex matrix is the limit of matrices with distinct eigenvalues. The singular value decomposition can be used to show that every complex matrix (square or not) is the limit of matrices with distinct singular values. This can be useful because of the partial uniqueness of the singular value decomposition in the case of distinct singular values.

7.3.6 **Corollary.** If $A \in M_{m,n}$ is given, and if $\|\cdot\|$ is a given norm on $M_{m,n}$, then for every $\epsilon > 0$ there exists a matrix $A_\epsilon \in M_{m,n}$ with distinct singular values such that $\|A - A_\epsilon\| < \epsilon$.

Proof: Suppose $m \leq n$. Let $A = V\Sigma W^*$ be a singular value decomposition of A, and let

$$\Sigma_\delta \equiv [\text{diag}(\sigma_1 + \delta, \sigma_2 + 2\delta, \dots, \sigma_m + m\delta) \mid 0]$$

with $0 \in M_{m,n-m}$. If all the singular values of A are equal, Σ_δ will have distinct diagonal entries for all $\delta > 0$. If not, and if $\delta > 0$ is chosen so that $m\delta$ is less than the smallest difference between successive distinct singular values, then Σ_δ will have distinct diagonal entries. In either event, $\Sigma_\delta \to \Sigma$ as $\delta \to 0$. If we set $A_\delta \equiv V\Sigma_\delta W^*$, then $\|A - A_\delta\|_2 = \|\Sigma - \Sigma_\delta\|_2 \to 0$ as $\delta \to 0$ since the Frobenius norm is unitarily invariant. But all norms on $M_{m,n}$ are equivalent, so we are done. The argument is similar if $m > n$. □

There is a simple transformation that permits one to convert results about eigenvalues of Hermitian matrices into results about singular values of arbitrary matrices.

7.3.7 **Theorem.** Let $A \in M_{m,n}$, let $q = \min\{m, n\}$, and define $\tilde{A} \in M_{m+n}$ by

$$\tilde{A} \equiv \begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} \tag{7.3.7a}$$

Let $\sigma_1, \sigma_2, \ldots, \sigma_q$ be nonnegative real numbers. The singular values of A are $\sigma_1, \sigma_2, \ldots, \sigma_q$ if and only if the $m+n$ eigenvalues of $\tilde{A}$ are $\sigma_1, \sigma_2, \ldots, \sigma_q, -\sigma_1, -\sigma_2, \ldots, -\sigma_q$, and $|m-n|$ additional 0's.

Proof: Suppose $m \geq n$ and let $A = V\Sigma W^*$ be a singular value decomposition of A. Write

$$\Sigma = \begin{bmatrix} S \\ 0 \end{bmatrix} \in M_{m,n}, \qquad 0 \in M_{m-n,n}$$

where $S = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$, and write the unitary factor $V \in M_m$ as $V = [V_1 \mid V_2]$, where $V_1 \in M_{m,n}$ and $V_2 \in M_{m,(m-n)}$. If we set $\hat{V} \equiv V_1/\sqrt{2}$ and $\hat{W} \equiv W/\sqrt{2}$, then the matrix

$$U \equiv \begin{bmatrix} \hat{V} & -\hat{V} & V_2 \\ \hat{W} & \hat{W} & 0 \end{bmatrix} \in M_{m+n}, \qquad 0 \in M_{n,m-n}$$

is unitary and one can verify by a direct calculation that

$$\tilde{A} = U \begin{bmatrix} S & 0 & 0 \\ 0 & -S & 0 \\ 0 & 0 & 0 \end{bmatrix} U^*$$

where the diagonal zero is an $(m-n)$-by-$(m-n)$ matrix. The argument is similar if $m < n$. □

Exercise. Let $A \in M_{m,n}$ be given. Show that the singular values of A^*, A^T, and $\bar{A}$ are the same as those of A. If $U \in M_m$ and $V \in M_n$ are unitary, show that the singular values of UAV are the same as those of A. If $c \in \mathbf{C}$, show that the singular values of cA are $|c|$ times the singular values of A.

As an immediate application of Theorem (7.3.7) we have perturbation results for singular values of arbitrary matrices that follow from the corresponding results for Hermitian matrices. They show that every matrix is perfectly conditioned with respect to singular value computations; they should be compared with (6.3.2) and (6.3.4) and the discussion of the condition number between. For a generalization of these results to arbitrary unitarily invariant norms see (7.4.51).

7.3.8 **Corollary.** Let $A, B \in M_{m,n}$, let $E \equiv B - A$, and let $q = \min\{m, n\}$. If $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_q$ are the singular values of A and $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_q$ are the singular values of B, then

(a) $|\sigma_i - \tau_i| \leq \|E\|_2$ for all $i = 1, 2, \ldots, q$; and
(b) $[\sum_{i=1}^{q} (\sigma_i - \tau_i)^2]^{1/2} \leq \|E\|_2$.

Proof: These two results are analogs of Weyl's inequality [(4.3.1); see also the exercise before (6.3.5)] and the Hoffman–Wielandt theorem for Hermitian matrices (6.3.8). They follow immediately from the stated results and (7.3.7). □

Exercise. Provide the details for the proof of (7.3.8). For (a), see Problem 36 of Section (5.6).

There is also an interlacing property for singular values; it follows from the interlacing property of eigenvalues of Hermitian matrices.

7.3.9 **Theorem.** Let $A \in M_{m,n}$ be a given matrix and let $\hat{A}$ be the matrix obtained by deleting any one column from A. Let $\{\sigma_i\}$ denote the singular values of A and let $\{\hat{\sigma}_i\}$ denote the singular values of $\hat{A}$, both arranged in nonincreasing order.

(a) If $m \geq n$, then

$$\sigma_1 \geq \hat{\sigma}_1 \geq \sigma_2 \geq \hat{\sigma}_2 \geq \cdots \geq \hat{\sigma}_{n-1} \geq \sigma_n \geq 0$$

(b) If $m < n$, then

$$\sigma_1 \geq \hat{\sigma}_1 \geq \sigma_2 \geq \hat{\sigma}_2 \geq \cdots \geq \sigma_m \geq \hat{\sigma}_m \geq 0$$

If a *row* of A is deleted instead of a column, the appropriate inequalities are obtained by interchanging m and n in (a) and (b).

Proof: The squares of the singular values of A are the eigenvalues of the Hermitian matrix $A^*A \in M_n$, and the squares of the singular values of $\hat{A}$ are the eigenvalues of $\hat{A}^*\hat{A} \in M_{n-1}$, which is a principal submatrix of A^*A if a column of A is deleted. The interlacing inequalities follow directly from the inclusion principle (4.3.15). If a row of A is deleted instead of a column, consider AA^* and $\hat{A}\hat{A}^*$ instead. □

As a final similarity between properties of eigenvalues of Hermitian matrices and properties of singular values, we have the following analog of the Courant–Fischer theorem (4.2.11).

7.3.10 Theorem. Let $A \in M_{m,n}$, let $q = \min\{m, n\}$, let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_q$ be the ordered singular values of A, and let k be a given integer with $1 \leq k \leq q$. Then

$$\min_{w_1, \ldots, w_{k-1} \in \mathbf{C}^n} \max_{\substack{x \neq 0,\, x \in \mathbf{C}^n \\ x \perp w_1, \ldots, w_{k-1}}} \frac{\|Ax\|_2}{\|x\|_2} = \sigma_k$$

and

$$\max_{w_1, \ldots, w_{n-k} \in \mathbf{C}^n} \min_{\substack{x \neq 0,\, x \in \mathbf{C}_n \\ x \perp w_1, \ldots, w_{n-k}}} \frac{\|Ax\|_2}{\|x\|_2} = \sigma_k$$

Proof: These formulae follow immediately from (4.2.12) and (4.2.13), since $\sigma_k^2(A)$ is an eigenvalue of A^*A. If $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k$ are the ordered eigenvalues of the Hermitian matrix A^*A, then $\sigma_k^2(A) = \lambda_{n-k+1}(A^*A)$, and (4.2.12) says that

$$\sigma_k^2(A) = \lambda_{n-k+1}(A^*A) = \min_{w_1, \ldots, w_{k-1} \in \mathbf{C}^n} \max_{\substack{x \neq 0,\, x \in \mathbf{C}^n \\ x \perp w_1, \ldots, w_{k-1}}} \frac{x^*A^*Ax}{x^*x}$$

$$= \min_{w_1, \ldots, w_{k-1} \in \mathbf{C}^n} \max_{\substack{x \neq 0,\, x \in \mathbf{C}^n \\ x \perp w_1, \ldots, w_{k-1}}} \left(\frac{\|Ax\|_2}{\|x\|_2}\right)^2$$

The second identity is proved in the same way. □

Problems

1. Let $P \in M_n$ be positive semidefinite. Show that P can be written as a polynomial in P^2, so if a given matrix U commutes with P^2, it must also commute with P. Use this to show that if $A \in M_n$ is normal, then its polar factors P and U commute.

2. Show that any $A \in M_n$ can be written as $A = Pe^{iH}$, where $P, H \in M_n$, P is positive semidefinite, and H is Hermitian. Show that H can be taken to be positive definite. To what extent are P and H determined by A? *Hint:* If $U \in M_n$ is unitary, and if $U = V\Lambda V^*$ is a unitary diagonalization of U, then $\Lambda = e^{iD}$, where D is a diagonal matrix with real main diagonal entries. What is e^{iVDV^*}?

3. Show that $A \in M_n$ has a zero singular value if and only if it has a zero eigenvalue.

4. Let $A \in M_{m,n}$ and let $q = \min\{m, n\}$. Show that the largest singular value of A is equal to the spectral norm of A. Show that the Frobenius norm of A satisfies the identity

$$\|A\|_2 = \left(\sum_{i=1}^{q} \sigma_i^2\right)^{1/2}$$

Show that $\sigma_1 \leq \|A\|_2 \leq \sqrt{n}\sigma_1$ and identify the cases of equality. Conclude that

$$|||A|||_2 \leq \|A\|_2 \leq \sqrt{n}\,|||A|||_2 \quad \text{for all} \quad A \in M_n \tag{7.3.11}$$

Show that these bounds are sharp by considering I and $\left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$.

5. If $k \leq \min\{m, n\}$ and v_k is the kth column of V and w_k is the kth column of W in a singular value decomposition (7.3.5) of A, show that

$$A^*v_k = \sigma_k w_k \quad \text{and} \quad Aw_k = \sigma_k v_k$$

where σ_k is the kth singular value of A. In particular, $v_k^* A w_k = \sigma_k$.

6. If one is given a large matrix A, how does one go about computing the rank of A numerically? Notice that the rank of A is equal to the number of nonzero singular values of A, so one way to compute the rank of A numerically is to normalize A to $A_1 \equiv A/\|A\|$ for some conveniently calculated norm $\|\cdot\|$, then compute a singular value decomposition for A_1 and take the rank of A to be the number of singular values of A_1 that are larger than some threshold. Why would you expect the numerical determination of the rank of A to be easier and more accurate if the ratio of the smallest to largest nonzero singular values of A is not near 0?

7. Let $A \in M_{m,n}$ have singular value decomposition $A = V\Sigma W^*$ and define $A^\dagger = W\Sigma^\dagger V^*$, where $\Sigma^\dagger$ is the transpose of Σ in which the positive singular values of A are replaced by their reciprocals. Show that

(a) $AA^\dagger$ and $A^\dagger A$ are Hermitian;
(b) $AA^\dagger A = A$; and
(c) $A^\dagger A A^\dagger = A^\dagger$.

Show that $A^\dagger = A^{-1}$ if A is square and nonsingular. The matrix $A^\dagger$ is called the *Moore–Penrose generalized inverse* of A. It exists for any matrix A, even for a singular square A and for a nonsquare A. Show further that $A^\dagger$ is uniquely determined by the above requirements (a)–(c).

8. A *least-squares solution* to the linear equations $Ax = b$ is a vector x such that $\|x\|_2$ is minimized among all vectors x for which $\|Ax - b\|_2$ is minimal. Show that $x = A^\dagger b$ is a least-squares solution to $Ax = b$.

9. Show that $A^{\dagger} = \lim_{t\to 0} A^*(AA^*+tI)^{-1}$, where $A^{\dagger}$ is defined in Problem 7.

10. The singular value decomposition (7.3.5) can be derived without explicit use of eigenvectors and eigenvalues. The (left and right) singular vectors and the singular values can be constructed directly from the variational characterization of the spectral norm. Consider $A \in M_n$ and the variational characterization (**) $\|A\|_2 = \max\{\|Ax\|_2 : \|x\|_2 = 1\}$. (a) Let $n \ge 2$ and let $B \in M_n$ have the special form

$$B = \begin{bmatrix} \sigma_1 & w^* \\ 0 & X \end{bmatrix}$$

where $\sigma_1 \equiv \|B\|_2$, $w \in \mathbf{C}^{n-1}$, and $X \in M_{n-1}$. Show that $w = 0$. *Hint:* If $\sigma_1 > 0$, consider $\zeta \equiv \left[\begin{smallmatrix} \sigma_1 \\ w \end{smallmatrix}\right]/(\sigma_1^2 + w^*w)^{1/2}$, show that $\|B\zeta\|_2^2 \ge \sigma_1^2 + w^*w$, and use (**). (b) Let $A \in M_n$, let $\sigma_1 \equiv \|A\|_2$, and use (**) to show that there is some unit vector x_1 such that $\|Ax_1\|_2 = \sigma_1$. Let $y_1 \equiv \sigma_1^{-1}Ax_1$. (c) Let $W_1, V_1 \in M_n$ be unitary matrices whose first columns are x_1 and y_1, respectively. Show that $V_1^*AW_1$ has spectral norm σ_1 and has the form of the matrix in (a). Conclude that $V_1^*AW_1 = \left[\begin{smallmatrix} \sigma_1 & 0 \\ 0 & X \end{smallmatrix}\right]$. (d) Formulate an induction procedure for deflating A by introducing additional columns and rows of off-diagonal zeroes by pre- and postmultiplication by unitary matrices, and obtain the singular value decomposition of A. (e) What if $A \in M_{m,n}$ is not square?

11. Let $A = V\Sigma W^*$ be a singular value decomposition of the matrix $A \in M_{m,n}$, suppose A has rank k, and let $q = \min\{m, n\}$. Show that the last $n-k$ columns of W form an orthonormal basis for the null space of A and that the first k columns of V form an orthonormal basis for the range of A.

12. Let $A \in M_{m,n}$ and $B \in M_{p,n}$. Show that an orthonormal basis for the intersection of the null spaces of A and B is given by the last several (how many?) columns of W, where $V\Sigma W'$ is a singular value decomposition of the partitioned matrix $\left[\begin{smallmatrix} A \\ B \end{smallmatrix}\right] \in M_{(m+p),n}$. *Hint:* When is $\left[\begin{smallmatrix} A \\ B \end{smallmatrix}\right]x = 0$ for $x \in \mathbf{C}^n$? How can you find an orthonormal basis for the intersection of the null spaces of k matrices $A_1, A_2, \ldots, A_k$, each having the same number of columns?

13. Show that the polar decomposition (7.3.2) and the singular value decomposition (7.3.5) are equivalent in the sense that each is easily derived from the other. *Hint:* Apply the spectral theorem to P.

14. Let $A \in M_n$. Show that A is diagonalizable if and only if there is a positive definite Hermitian matrix P such that $P^{-1}AP$ is normal. *Hint:* If $A = S\Lambda S^{-1}$, apply the polar decomposition (7.3.3) to S.

15. Use the singular value decomposition (7.3.5) (especially the statements about uniqueness of the decomposition) and Corollary (7.3.6) to prove Takagi's representation (4.4.4) for a complex symmetric matrix. *Hint:* If $A = A^T \in M_n$ has distinct singular values and if $A = V\Sigma W^*$, then $A = A^T = \bar{W}\Sigma V^T$. But then there exists a diagonal unitary matrix $D = \operatorname{diag}(e^{i\theta_1}, \ldots, e^{i\theta_n})$ such that $\bar{W} = VD$, so $A = V\Sigma W^* = V\Sigma(\bar{V}\bar{D})^* = V\Sigma DV^T = (VD^{1/2})\Sigma(VD^{1/2})^T \equiv U\Sigma U^T$. For the general case, use (7.3.6) and the selection principle (2.1.8) to perturb and pass to the limit.

16. Let $A, B \in M_{m,n}$, let $q = \min\{m, n\}$, let the ordered singular values of A be $\sigma_1(A) \ge \cdots \ge \sigma_q(A) \ge 0$, and similarly for B and $A+B$. Let $\tilde{A}, \tilde{B}, \tilde{A}+\tilde{B} \in M_{m+n}$ be the Hermitian matrices defined as in (7.3.7a). Show that $\sigma_k(A) = \lambda_{m+n-k+1}(\tilde{A})$ for $k = 1, 2, \ldots, q$ and similarly for B and $A+B$. Be careful: The singular values are arranged in decreasing order and the eigenvalues of the Hermitian matrix $\tilde{A}$ are arranged in increasing order. Use this identity and Weyl's theorem (4.3.7) to show that

$$\sigma_{i+j-1}(A+B) \le \sigma_i(A) + \sigma_j(B), \quad 1 \le i, j \le q \text{ and } i+j \le q+1$$

In particular, $\sigma_1(A+B) \le \sigma_1(A) + \sigma_1(B)$ (why is this not surprising?), and $\sigma_q(A+B) \le \min\{\sigma_q(A) + \sigma_1(B), \sigma_1(A) + \sigma_q(B)\}$.

17. Consider $A = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$ and $B = \left[\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ to show that the inequality $\sigma_i(A+B) \le \sigma_i(A) + \sigma_i(B)$ is *not* true for all $i = 1, 2, \ldots$, where $\{\sigma_i(A)\}$ and $\{\sigma_i(B)\}$ are the singular values of A and B, respectively, both arranged in decreasing order.

18. Let $A, B \in M_{m,n}$ be given, let $q = \min\{m, n\}$, let the ordered singular values of A be $\sigma_1(A) \ge \cdots \ge \sigma_q(A) \ge 0$ and similarly for B and $AB^* \in M_m$. Show that

$$\sigma_{i+j-1}(AB^*) \le \sigma_i(A)\sigma_j(B), \qquad 1 \le i, j \le q, \qquad i+j \le q+1$$

These inequalities may be thought of as a multiplicative analog of the additive inequalities in Problem 16 as well as a generalization of the submultiplicative property of the spectral norm when $m = n$. Why? *Hint:* Let $AB^* = WQ$ be a left polar decomposition of AB^*, with unitary $W \in M_m$ and positive semidefinite $Q \in M_m$. Show that $(x^*Qx)^2 = (x^*W^*AB^*x)^2 = [(A^*Wx)^*(B^*x)]^2 \le \|A^*Wx\|_2^2\|B^*x\|_2^2 = [(Wx)^*AA^*(Wx)](x^*BB^*x)$ for any $x \in \mathbf{C}^m$. Let $z_1, \ldots, z_{i-1}$ be orthonormal eigenvectors of AA^* corre-

sponding to the $i-1$ largest eigenvalues $\sigma_1^2(A), \ldots, \sigma_{i-1}^2(A)$ of AA^*, let $y_1, \ldots, y_{j-1}$ be orthonormal eigenvectors of BB^* corresponding to the $j-1$ largest eigenvalues $\sigma_1^2(B), \ldots, \sigma_{j-1}^2(B)$ of BB^*, and let $x_1 = W^*z_1$, $x_2 = W^*z_2, \ldots, x_{i-1} = W^*z_{i-1}, x_i = y_1, x_{i+1} = y_2, \ldots, x_{i+j-2} = y_{j-1}$. If x is orthogonal to x_k for $k = 1, 2, \ldots, i+j-2$, then both $(Wx)^*AA^*(Wx) \leq \sigma_i^2(A)\|x\|_2^2$ and $x^*BB^*x \leq \sigma_j^2(B)\|x\|_2^2$, so under these constraints we have $(x^*Qx)^2 \leq \sigma_i^2(A)\sigma_j^2(B)\|x\|_2^4$. Now invoke the Courant–Fischer theorem (4.2.11) to conclude that

$$\sigma_{i+j-1}^2(AB^*) = (\lambda_{n-1-j+2}([(AB^*)^*(AB^*)]^{1/2}))^2 \leq \sigma_i^2(A)\sigma_j^2(B)$$

19. Although the eigenvalues of AB and BA are always the same if $A, B \in M_n$, consider the examples $\left[\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ to show that the singular values of AB and BA need not be the same. Show, however, that the singular values of AB and B^*A^* are always the same.

20. Let X be an n-dimensional random vector whose components have zero means and finite variances. Let $\Sigma \equiv \text{Cov}(X) = E(XX^*)$ [see (4.5.3*)], assume that Σ is nonsingular, let $P = \Sigma^{1/2}$, and let $A, B \in M_n$ be given. The random vectors AX and BX have the same (zero) mean vectors, but there is no reason to expect that they have the same covariance matrices. Show that $\text{Cov}(AX) = \text{Cov}(BX)$ if and only if $A = B(PUP^{-1})$ for some unitary matrix $U \in M_n$. *Hint:* If $A\Sigma A^* = B\Sigma B^*$, then $(AP)(AP)^* = (BP)(BP)^*$. If RW is a polar decomposition of BP, show that RV is a polar decomposition of AP for some unitary $W, V \in M_n$. What is R? Conclude that $A = B(PW^*VP^{-1}) = B(PUP^{-1})$. To what extent is U determined? What if $\Sigma = I$? What if $B = I$?

21. Consider the matrix $A_\epsilon \in M_n$ given by

$$A_\epsilon = \begin{bmatrix} 0 & 1 & & 0 \\ \vdots & & \ddots & \\ 0 & & \ddots & 1 \\ \epsilon & \cdot\ 0 & \cdot\cdot & 0 \end{bmatrix}, \qquad \epsilon > 0$$

Show that the characteristic polynomial of A_ϵ is $t^n - \epsilon$. *Hint:* Compute $\det(tI - A_\epsilon)$ by a Laplace cofactor expansion along the first column. Show that the eigenvalues of A_ϵ are the n choices of $\sqrt[n]{\epsilon}$. Show that the singular values of A_ϵ are 1, with multiplicity $n-1$, and ϵ. Now let $n = 10$, $\epsilon = 10^{-10}$ and observe that the perturbation $A_0 \to A_\epsilon$ results in a .1 perturbation of the eigenvalues of A_0, but only a 10^{-10} perturbation of any singular value of A_0. What is the spectral condition number of A_ϵ? This is an example of the assertion following Theorem (7.3.7) that every

matrix is well conditioned with respect to singular value computations, whereas a given matrix may be poorly conditioned with respect to eigenvalue computations.

22. Let $A = [a_{ij}] \in M_n$ be given. Show that if A has a "small" row or column, then it must also have a "small" singular value. More precisely, let $A = [r_1\, r_2 \ldots r_n]^T$, where $r_i \in \mathbf{C}^n$ and r_i^T is the ith row of A. Place the set of Euclidean norms of the rows $\{\|r_i\|_2 : i = 1, \ldots, n\}$ in increasing order and denote the resulting ordered values by $R_1 \leq R_2 \leq \cdots \leq R_n$. Show that

$$\sum_{i=1}^{k} \sigma_{n-i+1}^2 \leq \sum_{i=1}^{k} R_i^2 \quad \text{for} \quad k = 1, 2, \ldots, n$$

with a similar upper bound involving the norms of the columns. Recall that the singular values are ordered with $\sigma_n \leq \sigma_{n-1} \leq \cdots \leq \sigma_1$. *Hint:* The squared singular values are the eigenvalues of the Hermitian matrix AA^*. What are the main diagonal entries of AA^*? Use majorization and Theorem (4.3.26). Consider A^*A for the column sum inequalities. Compare with Problem 19 of Section (4.3).

23. There is a natural analog of the singular value decomposition in which the unitary factors are replaced by complex orthogonal factors. Unlike the singular value decomposition, however, this factorization cannot always be achieved; recall from Problem 7 of Section (2.3) that the orthogonal analog of the Schur unitary upper triangular factorization cannot always be achieved either. If $A \in M_{m,n}$ can be written in the form $A = P\Lambda Q^T$, where $P \in M_m$ and $Q \in M_n$ are complex orthogonal and $\Lambda = [\lambda_{ij}] \in M_{m,n}$ is "diagonal" in the sense that $\lambda_{ij} = 0$ if $i \neq j$, show that $AA^T \in M_m$ is diagonalizable and $\text{rank}\, A = \text{rank}\, AA^T$. These two conditions are also sufficient to ensure the existence of the indicated factorization $A = P\Lambda Q^T$. What does this say if A is real? Give an example of an $A \in M_2$ that cannot be written as $A = P\Lambda Q^T$ with complex orthogonal $P, Q \in M_2$ and diagonal $\Lambda \in M_2$.

24. Explain why the singular value decomposition may be thought of as a generalization of the spectral theorem for normal matrices.

25. Theorem (2.5.5) on simultaneous unitary diagonalization of a family of normal matrices has an analog for the singular value decomposition. Let $\mathfrak{F} = \{A_i : i \in \mathfrak{I}\} \subset M_{m,n}$ and suppose there are unitary matrices $V \in M_m$ and $W \in M_n$ such that every V^*A_iW is "diagonal" in the sense of Problem 23; that is, its i, j entry is 0 if $i \neq j$. Show that (a) Each $A_i^*A_j \in M_n$ is normal and $\mathcal{G} = \{A_iA_j^* : i, j \in \mathfrak{I}\} \subset M_m$ is a commuting family.

(b) $A_iA_j^*A_k = A_kA_j^*A_i$ for every $i, j, k \in \mathcal{I}$. Each of these necessary conditions is also sufficient for the family $\mathfrak{F}$ to have a simultaneous factorization of the form of the singular value decomposition.

26. Finding a simultaneous factorization of two given matrices $A, B \in M_{m,n}$ of the form of the singular value decomposition is an interesting special case of the preceding problem. Show that there are unitary matrices $V \in M_m$, $W \in M_n$ such that $A = V\Sigma W^*$ and $B = V\Lambda W^*$ with $\Sigma, \Lambda \in M_{m,n}$ "diagonal" if and only if AB^* and B^*A are both normal. *Hint:* To show that the condition is sufficient, argue that it suffices to consider only the case in which $A = \Sigma$ is nonnegative and "diagonal." If the equal diagonal entries of Σ are grouped together, show that if ΣB^* and $B^*\Sigma$ are normal, then B is a partitioned block diagonal matrix with all but perhaps one (if A is singular) normal block. For each block, use either the spectral theorem for normal matrices or the singular value decomposition to obtain the conclusion.

27. If we wish to have unitary matrices $V \in M_m$ and $W \in M_n$ such that every member of the family $\mathfrak{F} = \{A_i : i \in \mathcal{I}\} \subset M_{m,n}$ can be written as $A_i = V\Sigma_i W^*$ with each Σ_i "diagonal," show that it is necessary, but not sufficient when there are three or more matrices in the family, that $A_iA_j^* \in M_m$ and $A_i^*A_j \in M_n$ be normal for every $i, j \in \mathcal{I}$. *Hint:* Consider the family

$$\mathfrak{F} = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & i \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \right\}$$

Explain what part of the proof in the case of two matrices does not work when there are more than two.

Further Readings and Notes. Sylvester proved the singular value decomposition for real square matrices in 1889. What seems to be the first proof of the singular value decomposition for general m-by-n complex matrices is in C. Eckart and G. Young, "A Principal Axis Transformation for Non-Hermitian Matrices," *Bull. Amer. Math. Soc.* 45 (1939), 118–121. Eckart and Young's paper also contains the result that two matrices $A, B \in M_{m,n}$ have a simultaneous factorization of the form of the singular value decomposition in which the respective "diagonal" factors are both *real* if and only if AB^* and B^*A are both Hermitian. For a survey of results and further references on simultaneous factorization of families in the form of the singular value decomposition, see P. M. Gibson, "Simultaneous Diagonalization of Rectangular Complex Matrices," *Linear Algebra Appl.* 9 (1974), 45–53.

7.4 Examples and applications of the singular value decomposition

There are many applications of the polar form and the singular value decomposition. Some are given in the problems, and several are discussed in the following examples.

7.4.1 **Example.** If $A \in M_n$ is a given invertible matrix, then all matrices that are sufficiently close to A (with respect to any norm) are invertible, too. In some statistical modeling problems it is required to find a "nearest singular matrix" to A in the sense of least squares; that is, we want to find a matrix B such that $A+B$ is singular and $\|B\|_2$ is as small as possible.

Let $|\!|\!|\cdot|\!|\!|$ be any matrix norm, and consider $A+B = A(I+A^{-1}B)$, which we assume to be singular. If $|\!|\!|A^{-1}B|\!|\!| < 1$, then $I+A^{-1}B$, and hence $A+B$, would be invertible by (5.6.16). Thus, $1 \leq |\!|\!|A^{-1}B|\!|\!| \leq |\!|\!|A^{-1}|\!|\!| \, |\!|\!|B|\!|\!|$, so if $A+B$ is singular and A is invertible, we must have $|\!|\!|B|\!|\!| \geq 1/|\!|\!|A^{-1}|\!|\!|$. If we choose for $|\!|\!|\cdot|\!|\!|$ the spectral norm, and if $A = V\Sigma W^*$ is a singular value decomposition of A, then $|\!|\!|A^{-1}|\!|\!|_2 = |\!|\!|W\Sigma^{-1}V^*|\!|\!|_2 = |\!|\!|\Sigma^{-1}|\!|\!|_2 = 1/\sigma_n$, where σ_n is the smallest singular value of A. Then, any B such that $A+B$ is singular must satisfy $|\!|\!|B|\!|\!|_2 \geq \sigma_n(A)$. But if we choose for B the matrix $B = VEW^*$, where $E \equiv \text{diag}(0, 0, \ldots, 0, -\sigma_n)$, then $|\!|\!|B|\!|\!|_2 = |\!|\!|E|\!|\!|_2 = \sigma_n = \|E\|_2 = \|B\|_2$ and $A+B$ is singular (and has rank $n-1$).

More generally, if we want to find a "nearest rank k matrix" to a given matrix A, singular or nonsingular, with respect to the Frobenius norm, we may choose $A+B$, where $B = VEW^*$ as before, but $E = \text{diag}(0, \ldots, 0, -\sigma_{k+1}, \ldots, -\sigma_n)$. See Problem 1 at the end of this section for a proof and Example (7.4.52) for a generalization of this result from the Frobenius norm to all unitarily invariant norms.

The case $k=1$ occurs frequently enough in the applications that it deserves special mention. A best least-squares approximation to a given matrix $A = V\Sigma W^* \in M_n$ by a rank 1 matrix $X \in M_n$ is $X = A+B = V(\Sigma+E)W^* = V\text{diag}(\sigma_1, 0, \ldots, 0)W^* = \sigma_1 vw^*$, where σ_1 is the largest singular value of A, and v and w are the first columns of the unitary matrices V and W in a singular value decomposition of A, respectively. A useful observation about v and w is that they are unit vector solutions of the pair of Hermitian eigenvalue-eigenvector problems

$$AA^*v = \sigma_1^2 v, \qquad A^*Aw = \sigma_1^2 w$$

where σ_1^2 is the largest eigenvalue of the positive semidefinite matrix A^*A (and AA^*). This observation does not uniquely determine v and w, of course; one difficulty is that the eigenspaces associated with σ_1^2 need not

be one-dimensional. If σ_1^2 is a simple eigenvalue of A^*A (and hence of AA^*), however, the eigenvectors v and w are determined up to scalar factors of modulus 1 and must therefore be scalar multiples of the respective first columns of the unitary matrices V and W in a singular value decomposition $A=V\Sigma W^*$. In this case, for fixed choices of unit eigenvectors v and w, a best rank 1 approximation to A must be of the form $e^{i\theta}\sigma_1 vw^*$ for some $\theta\in\mathbf{R}$. The scalar factor $e^{i\theta}$ must be chosen to minimize $\|A-e^{i\theta}\sigma_1 vw^*\|_2^2=\|A\|_2^2-2\sigma_1\operatorname{Re}[\operatorname{tr} e^{-i\theta}A(vw^*)^*]+\sigma_1^2\|v\|_2^2\|w\|_2^2$, a problem equivalent to maximizing $\operatorname{Re}[\operatorname{tr} e^{-i\theta}A(vw^*)^*]=\operatorname{Re}[e^{-i\theta}v^*Aw]$. But $Aw=V\Sigma W^*w=e^{i\phi}\sigma_1 v$ for some $\phi\in\mathbf{R}$ [see Problem 5 of Section (7.3)], and hence $|v^*Aw|=\sigma_1>0$. Thus, the optimal scalar factor is $e^{i\theta}=v^*Aw/|v^*Aw|=v^*Aw/\sigma_1$ and a best rank 1 approximation to A is

$$e^{i\theta}\sigma_1 vw^*=(v^*Aw)vw^*$$

This shows that if the largest eigenvalue of A^*A is simple, a best rank 1 least-squares approximation to A can be constructed without further effort from the solutions of two Hermitian eigenvalue problems. The condition of a simple maximal eigenvalue for A^*A is met, for example, by any nonnegative matrix $A\in M_n(\mathbf{R})$ such that AA^T is positive or, more generally, irreducible [see Problem 17 in Section (8.4)].

7.4.2 **Example.** In Theorem (5.7.17) we showed that a vector norm $G(\bullet)$ on M_n satisfies the condition

$$G(A_1)G(A_2)\cdots G(A_k)\ge\rho(A_1\cdots A_k)$$

for all $A_1,A_2,\ldots,A_k\in M_n$ and all $k=1,2,\ldots$ if and only if $G(\bullet)$ has a compatible vector norm on $\mathbf{C}^n$. A crucial step in this argument was to show that if $G(\bullet)$ obeys this inequality with respect to the spectral radius, then there is some finite constant $c>0$ such that $G(A_1)G(A_2)\cdots G(A_k)\ge c\|A_1A_2\cdots A_k\|_2$, and the key to showing this is the singular value decomposition of the product $A_1A_2\cdots A_k$. The details are in Lemma (5.7.16).

7.4.3 **Example.** Suppose one wishes to solve a system of linear equations $Ax=b$, where $A\in M_{m,n}$ and $b\in\mathbf{C}^m$ are given, and A has rank k. If $A=V\Sigma W^*$ is a singular value decomposition of A, then $V\Sigma W^*x=b$, or

$$\Sigma(W^*x)=V^*b \tag{7.4.4}$$

If $m>k$, then the last $m-k$ rows of Σ are 0, and hence if there is to be a solution in this case it is necessary (and also sufficient) that the last $m-k$ entries of V^*b be zero. Thus, the system $Ax=b$ is solvable when $m>k$ if

and only if b is orthogonal to the last $m-k$ left singular vectors of A. If b satisfies this consistency condition, and if $V=[v_1\ldots v_m]$ and $W=[w_1\ldots w_n]$, then (7.4.4) says that

$$(W^*x)^*=\left[\frac{b^*v_1}{\sigma_1},\ldots,\frac{b^*v_k}{\sigma_k},0,\ldots,0\right]^*$$

and hence the vector

$$x=\sum_{i=1}^{k}\frac{v_i^*b}{\sigma_i}w_i \tag{7.4.5}$$

is a solution. Since $Aw_j=V(\Sigma W^*w_j)=0$ for all $j>k$, any linear combination of the last $n-k$ right singular vectors of A (if any) is in the null space of A and hence the vector

$$x=\sum_{i=1}^{k}\frac{v_i^*b}{\sigma_i}w_i+\sum_{i=k+1}^{n}c_iw_i$$

will be a solution to $Ax=b$ for any $c_{k+1},\ldots,c_n\in\mathbf{C}$; this last sum is, of course, absent if $n=k$. Because the vectors $\{w_i\}$ are orthonormal, the solution with minimum l_2 norm is obtained when all $c_i=0$. Notice that the last $m-k$ left singular vectors of A span the null space of AA^*, which is the same as the null space of A^*, so requiring that b be orthogonal to the last $m-k$ left singular vectors of A is the same as requiring that b be orthogonal to every solution to $A^*x=0$.

Exercise. If not all of the last $m-k$ elements of V^*b are zero, then the system $Ax=b$ is inconsistent and there is no solution at all. For some purposes, one may be satisfied with a "least-squares" solution, however, which is a vector $x\in\mathbf{C}^n$ of minimum l_2 norm such that $\|Ax-b\|_2$ is minimized. Show that (7.4.5) gives such a least-squares solution.

7.4.6 **Example.** What is the best least-squares approximation to a given $A\in M_n$ by a scalar multiple of a unitary matrix? Recall that the l_2 norm on M_n is generated by the inner product $[A,B]\equiv\operatorname{tr}AB^*$, and that if U is unitary, then

$$\|U\|_2^2=[U,U]=\operatorname{tr}UU^*=\operatorname{tr}I=n$$

For any $c\in\mathbf{C}$ and for any unitary $U\in M_n$ we have

$$\|A-cU\|_2^2=[A-cU,A-cU]=\|A\|_2^2-2\operatorname{Re}\{\bar{c}[A,U]\}+n|c|^2$$

which is minimized when $c=[A,U]/n$, and hence

$$\|A-cU\|_2^2\ge\|A\|_2^2-\frac{1}{n}|[A,U]|^2$$

If we define

$$u(A) \equiv \max_{\text{unitary } U\in M_n} |[A,U]| \qquad (7.4.7)$$

then we obtain a quantity analogous to the numerical radius $r(A)$, for which the maximum of the inner product is taken not over unitary matrices but over all rank 1 Hermitian matrices of Frobenius norm 1. Unlike the numerical radius, however, the function $u(A)$ is a matrix norm on M_n [see Problem 5 and Example (7.4.54)].

It is easy to identify the value of $u(A)$ as well as the extremal unitary matrix. Let a singular value decomposition of A be $A=V\Sigma W^*$. Then

$$\begin{aligned}
u(A) &= \max_{\text{unitary } U} |[A,U]| = \max_{\text{unitary } U} |[V\Sigma W^*, U]| \\
&= \max_{\text{unitary } U} |\operatorname{tr} V\Sigma W^* U^*| = \max_{\text{unitary } U} |\operatorname{tr} \Sigma(W^*U^*V)| \\
&= \max_{\text{unitary } U} |\operatorname{tr} \Sigma U| = \max_{\text{unitary } U=[u_{ij}]} \left|\sum_{i=1}^{n} \sigma_i u_{ii}\right| \\
&\le \max_{\text{unitary } U=[u_{ij}]} \sum_{i=1}^{n} \sigma_i |u_{ii}| \le \sum_{i=1}^{n} \sigma_i
\end{aligned}$$

But if $A=PU$ is the polar form of A, then

$$[A,U]=\operatorname{tr} PUU^* = \operatorname{tr} P = \sum_{i=1}^{n} \sigma_i$$

so the upper bound given for $u(A)$ is sharp, $u(A)=\sigma_1(A)+\cdots+\sigma_n(A)$, and a best least-squares approximation of A by a multiple of a unitary matrix is given by

$$\frac{1}{n}(\sigma_1+\cdots+\sigma_n)U$$

if $A=PU$ is the polar form of A and $\sigma_1,\ldots,\sigma_n$ are its singular values. If a singular value decomposition $A=V\Sigma W^*$ is given, then $U=VW^*$. The error in the approximation is

$$\left\|A-\frac{u(A)}{n}U\right\|_2^2 = \|A\|_2^2 - \frac{1}{n}|[A,U]|^2 = \sum_{i=1}^{n}\sigma_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}\sigma_i\right)^2$$

which is 0 only when the Cauchy–Schwarz inequality

$$\left(\sum_{i=1}^{n}\sigma_i 1\right)^2 \le \left(\sum_{i=1}^{n} 1^2\right)\left(\sum_{i=1}^{n}\sigma_i^2\right)$$

is an inequality. Thus, A can be perfectly approximated by a multiple of a unitary matrix only when all its singular values are equal.

7.4.8 **Example.** Suppose $A,B\in M_{m,n}$ are given, and we wish to determine whether A was produced by "rotating" B; that is, is $A=UB$ for some unitary matrix $U\in M_m$? More generally, if we consider all the possible "rotations" UB of the given matrix B, how well can we approximate A in the sense of least squares? This is known in factor analysis as the problem of finding a "procrustean transformation" of B.

The computations are very similar to those in the previous example; we seek to choose U to minimize $\|A-UB\|_2$, and we compute, as before,

$$\|A-UB\|_2^2 = [A-UB, A-UB] = \|A\|_2^2 - 2\operatorname{Re}[A,UB] + \|B\|_2^2$$

Thus, we must find a unitary matrix U that maximizes $\operatorname{Re}[A,UB] = \operatorname{Re}\operatorname{tr} AB^*U^*$. If $AB^*=V\Sigma W^*$ is a singular value decomposition of AB^*, then

$$\begin{aligned}
\operatorname{Re}\operatorname{tr} AB^*U^* &= \operatorname{Re}\operatorname{tr} V\Sigma W^*U^* = \operatorname{Re}\operatorname{tr}\Sigma W^*U^*V \\
&= \operatorname{Re}\sum_{i=1}^{m}\sigma_i(AB^*)t_{ii}
\end{aligned}$$

where $T=[t_{ij}]=W^*U^*V$ is a unitary matrix. This sum is maximized when all $t_{ii}=1$, that is, when $U=VW^*$; VW^* is just the unitary part of a polar decomposition of AB^*.

Thus, a best least-squares approximation to $A\in M_{m,n}$ by a matrix of the form UB, where $B\in M_{m,n}$ and $U\in M_m$ is unitary, is given by $UB=(VW^*)B$, where $AB^*=V\Sigma W^*$ is a singular value decomposition of AB^*, or $AB^*=P(VW^*)$ is a polar decomposition of AB^*; we do not need to know V and W separately. The error in this approximation is given by

$$\begin{aligned}
\min\{\|A-UB\|_2 : U\in M_m \text{ is unitary}\} &= \|A-(VW^*)B\|_2 \\
&= \left[\|A\|_2^2+\|B\|_2^2-2\sum_{i=1}^{m}\sigma_i(AB^*)\right]^{1/2}
\end{aligned}$$

where $\{\sigma_i(AB^*)\}$ is the set of singular values of AB^*.

If we want to know whether A is exactly a rotation of B, then an obvious necessary condition is that $\|A\|_2=\|B\|_2$, and a necessary and sufficient condition is that

$$\|A\|_2^2 = \|B\|_2^2 = \sum_{i=1}^{m}\sigma_i(AB^*)$$

where $\{\sigma_i(AB^*)\}$ is the set of singular values of AB^*.

Finally, if we consider the special case $m=n$ and $B=I$, then we have the result that a best least-squares approximation of a given matrix $A\in M_n$ by a unitary matrix $U\in M_n$ is given by $U=VW^*$, where $A=V\Sigma W^*$ is

a singular value decomposition of A, or where $A=PU=P(VW^*)$ is a polar decomposition of A; the error in the approximation is

$$\|A-VW^*\|_2^2=\|A\|_2^2+\|I\|_2^2-2\sum_{i=1}^{n}\sigma_i(A)$$

$$=\sum_{i=1}^{n}\sigma_i^2(A)+n-2\sum_{i=1}^{n}\sigma_i(A)=\sum_{i=1}^{n}(\sigma_i(A)-1)^2$$

where $\{\sigma_i(A)\}$ is the set of singular values of A.

As part of the discussion in the preceding example, we found a solution to the problem of maximizing $\text{Re tr } AU$ over all unitary matrices U. For convenience of later reference, we summarize this result as

7.4.9 **Theorem.** Let $A\in M_n$ be a given matrix, and let $A=V\Sigma W^*$ be a singular value decomposition of A. Then (a) The problem

$$\max\{\text{Re tr } AU: U\in M_n \text{ is unitary}\}$$

has the solution $U=WV^*$, and the value of the maximum is $\sigma_1(A)+\cdots+\sigma_n(A)$, where $\{\sigma_i(A)\}$ is the set of singular values of A. (b) There exists a unitary matrix $U\in M_n$ such that $AU\in M_n$ is a positive semidefinite Hermitian matrix. A unitary matrix U is a maximizing matrix for the problem in (a) if and only if AU is positive semidefinite; U is uniquely determined if A is nonsingular. The eigenvalues of AU are the singular values of A.

Proof: Compute

$$\text{Re tr } AU=\text{Re tr } V\Sigma W^*U=\text{Re tr } \Sigma(W^*UV)=\sum_{i=1}^{n}\text{Re } \sigma_i(W^*UV)_{ii}$$

which is maximized only when all $(W^*UV)_{ii}=1$. Since W^*UV is unitary, this happens if and only if $W^*UV=I$, or $U=WV^*$. For this choice of U, $AU=V\Sigma W^*WV^*=V\Sigma V^*$, which is Hermitian and positive semidefinite since $\Sigma=\text{diag}(\sigma_1,\ldots,\sigma_n)$ and all $\sigma_i\geq 0$. If $U_1\in M_n$ is any unitary matrix for which AU_1 is positive semidefinite, the eigenvalues of AU_1 are the singular values of A because the singular values are unitarily invariant. The uniqueness in the nonsingular case follows from the uniqueness part of (7.3.3). □

For any matrix $A\in M_{m,n}$, AA^* and A^*A are both positive semidefinite, and $\text{tr } AA^*=\text{tr } A^*A=\sigma_1^2(A)+\cdots+\sigma_{\min\{m,n\}}^2(A)$, which may be thought of as a sum of products of singular values of A and A^*, respectively, since the singular values of A^* are the same as the singular values of A. This simple observation has a generalization to any pair of matrices A, B for which the products AB and BA are defined and positive semidefinite. This result is useful in approaching several types of matrix optimization problems.

7.4.10 **Theorem.** Let $A\in M_{m,n}$, $B\in M_{n,m}$, and $q=\min\{m,n\}$. Let $\sigma_1(A),\ldots,\sigma_q(A)$ and $\sigma_1(B),\ldots,\sigma_q(B)$ denote the singular values of A and B, respectively, arranged in nonincreasing order. If both $AB\in M_m$ and $BA\in M_n$ are positive semidefinite, then there exists a permutation τ of the integers $1,2,\ldots,q$ such that

$$\text{tr } AB=\text{tr } BA=\sum_{i=1}^{q}\sigma_i(A)\sigma_{\tau(i)}(B) \tag{7.4.11}$$

Proof: If $m=n$, if each of A and B is positive semidefinite, and if A and B commute, then they can be simultaneously unitarily diagonalized as $A=U\Lambda U^*$ and $B=UMU^*$, where $U\in M_m$ is unitary, $\Lambda=\text{diag}(\lambda_1,\ldots,\lambda_m)$, $M=\text{diag}(\mu_1,\ldots,\mu_m)$, and all λ_i,μ_i are nonnegative. In this case we have

$$\text{tr } AB=\text{tr }(U\Lambda U^*)(UMU^*)=\text{tr } U\Lambda MU^*=\text{tr } \Lambda M=\sum_{i=1}^{m}\lambda_i\mu_i$$

Since the eigenvalues λ_i,μ_i are also the singular values of A and B, the theorem is proved in this particular case.

There is no loss of generality to assume that $m\leq n$, for if $m>n$, one can just interchange A and B in the statement of the theorem.

To prove the theorem in general, we claim that it suffices to show that for any pair of matrices $A\in M_{m,n}$ and $B\in M_{n,m}$ such that $m\leq n$ and both AB and BA are positive semidefinite, there is a unitary matrix $V\in M_n$ and a matrix $Y\in M_{m,n}$ with orthonormal rows such that the transformation

$$\hat{A}=Y^*AV \quad\text{and}\quad \hat{B}=V^*BY \tag{7.4.12}$$

produces a pair of commuting positive semidefinite n-by-n matrices $\hat{A}$ and $\hat{B}$. In this event we have

$$\text{tr } AB=\text{tr } ABYY^*=\text{tr } Y^*ABY=\text{tr }(Y^*AV)(V^*BY)$$

$$=\sum_{i=1}^{m}\sigma_i(Y^*AV)\sigma_{\tau(i)}(V^*BY)=\sum_{i=1}^{m}\sigma_i(\hat{A})\sigma_{\tau(i)}(\hat{B})$$

by the above observation. Notice that $\hat{A}^*\hat{A}=V^*A^*YY^*AV=V^*A^*AV=(AV)^*(AV)$, so the singular values of $\hat{A}$ are the same as those of AV, which are the same as those of A since $(AV)(AV)^*=AA^*$. A similar argument shows that the singular values of $\hat{B}$ are the same as those of B, so we conclude that

$$\text{tr } AB=\sum_{i=1}^{m}\sigma_i(A)\sigma_{\tau(i)}(B)$$

as claimed. We proceed in three steps to establish the existence of a transformation of the form (7.4.12) with the required properties.

(1) Let A and B satisfy the hypotheses of the theorem. Recall from (1.3.20) that the eigenvalues of BA are the same as those of AB (counting multiplicities) together with an additional $n-m$ zero eigenvalues. If $\lambda_1,\ldots,\lambda_m$ are the eigenvalues of AB and $\Lambda=\operatorname{diag}(\lambda_1,\ldots,\lambda_m)$, then since both AB and BA are Hermitian by assumption, there are unitary matrices $U\in M_m$ and $V\in M_n$ such that

$$AB=U\Lambda U^* \quad\text{and}\quad BA=V\begin{bmatrix}\Lambda & 0\\ 0 & 0\end{bmatrix}V^*$$

If we partition $V=[V_1\,|\,V_2]$ with $V_1\in M_{n,m}$ and $V_2\in M_{n,n-m}$, then V_1 is a matrix with orthonormal columns, so $V_1^*V_1=I\in M_m$. Then $\Lambda=U^*ABU$ and $BA=V_1\Lambda V_1^*$, so $BA=(V_1U^*)AB(UV_1^*)$. Let $Y=UV_1^*\in M_{m,n}$ and observe that $YY^*=UV_1^*V_1U^*=UU^*=I$, so Y has orthonormal rows and $BA=Y^*ABY$. Set $\hat{A}\equiv Y^*A\in M_n$ and $\hat{B}\equiv BY\in M_n$, and compute $\hat{A}\hat{B}=Y^*ABY=BA$ and $\hat{B}\hat{A}=BYY^*A=BA$; the product BA is positive semidefinite by assumption. Thus, there is a transformation of the form (7.4.12) [with $V=I$] that yields a commuting pair of n-by-n matrices whose product is positive semidefinite. The individual terms $\hat{A}$ and $\hat{B}$ may not be positive semidefinite, however; a further transformation of the form (7.4.12) may be required to achieve this.

(2) Without loss of generality we may now assume that $m=n$, that $A,B\in M_n$ commute, and that the product AB is positive semidefinite. If $(AB)x=\lambda x$ with $x\neq 0$, then $(AB)(Ax)=ABAx=AABx=A(ABx)=A\lambda x=\lambda(Ax)$, so each of the eigenspaces of the Hermitian matrix AB is invariant under A. The same argument shows that each of these eigenspaces is also invariant under B. Thus, if $U=[u_1\ldots u_n]$ is a unitary matrix whose columns consist of eigenvectors of AB, and if the columns are arranged so that all the eigenvectors corresponding to the same eigenvalue of AB occur contiguously, then both U^*AU and U^*BU must be block diagonal with

$$\hat{A}=U^*AU=\operatorname{diag}(A_1,A_2,\ldots,A_r),\qquad \hat{B}=U^*BU=\operatorname{diag}(B_1,B_2,\ldots,B_r)$$

where $A_i,B_i\in M_{k_i}$, $1\leq k_i\leq n$, $k_1+\cdots+k_r=n$, and each $A_iB_i=B_iA_i=\lambda_iI\in M_{k_i}$, where $\lambda_1,\lambda_2,\ldots,\lambda_r$ are the *distinct* (nonnegative) eigenvalues of the positive semidefinite matrix AB.

(3) Without loss of generality we may now assume that $m=n$, $A,B\in M_n$, A and B commute, and $AB=\lambda I$ with $\lambda\geq 0$. If $\lambda>0$, then both A and B are nonsingular and $B=\lambda A^{-1}$. Use (7.4.9) to find a unitary matrix $U\in M_n$ such that $\hat{A}\equiv AU$ is positive semidefinite. But then $\hat{B}\equiv U^*B=\lambda U^*A^{-1}=\lambda(AU)^{-1}$ is also positive semidefinite since $\lambda>0$ and $(AU)^{-1}$ is positive semidefinite. Furthermore, $(U^*B)(AU)=U^*\lambda IU=\lambda I=AB=(AU)(U^*B)$, so $\hat{A}$ and $\hat{B}$ commute. This is a transformation of the form (7.4.12), so we are done if $\lambda>0$.

If $\lambda=0$, then $AB=BA=0$. Again choose a unitary U such that AU is positive semidefinite. Then $0=AB=(AU)(U^*B)=(U^*B)(AU)=U^*0U=0$, so AU and U^*B commute and every eigenspace of the Hermitian matrix AU is invariant under U^*B. If $W=[w_1\ldots w_n]$ is a unitary matrix whose columns consist of eigenvectors of AU, and if the columns are arranged so that all the eigenvectors corresponding to the same eigenvalue of AU occur contiguously, then both $W^*(AU)W$ and $W^*(U^*B)W$ are block diagonal with

$$W^*(AU)W=\operatorname{diag}(\Lambda_1,\ldots,\Lambda_r),\qquad W^*(U^*B)W=\operatorname{diag}(B_1,\ldots,B_r)$$

Λ_i and B_i are the same size, and $\Lambda_i=\lambda_iI$, $i=1,2,\ldots,r$, where $\lambda_1,\lambda_2,\ldots,\lambda_r$ are the *distinct* (nonnegative) eigenvalues of the positive semidefinite matrix AU. We have $\Lambda_iB_i=B_i\Lambda_i=0$ for all $i=1,\ldots,r$. If $\lambda_i\neq 0$, then the pair $\Lambda_i=\lambda_iI$, $B_i=0$ is a commuting pair of positive semidefinite matrices, as required. If $\lambda_i=0$, then B_i is not necessarily zero, but there is a unitary matrix U_i such that $U_i^*B_i$ is positive semidefinite [apply (7.4.9) to B^*] and, in this case, $\Lambda_iU_i=0$ and $U_i^*B_i$ constitute a commuting positive semidefinite pair obtained by a transformation of the form (7.4.12). This completes examination of all possible cases. □

7.4.13 **Example.** As a variation on the rotation problem in (7.4.8), let $A,B\in M_{m,n}$ be given and suppose we wish to determine whether A was produced by a two-sided "rotation" of B; that is, is $A=UBV$ for some unitary matrices $U\in M_m$, $V\in M_n$? More generally, if we consider all the possible two-sided "rotations" UBV of the given matrix B, how well can we approximate A in the sense of least squares?

As before, we seek to choose unitary matrices $U\in M_m$ and $V\in M_n$ to minimize $\|A-UBV\|_2$, and we compute as before

$$\|A-UBV\|_2^2=[A-UBV,A-UBV]=\|A\|_2^2-2\operatorname{Re}[A,UBV]+\|B\|_2^2$$

Thus, we must find unitary matrices $U\in M_m$ and $V\in M_n$ that maximize $\operatorname{Re}[A,UBV]=\operatorname{Re}\operatorname{tr}AV^*B^*U^*$. Maximizing unitary matrices U_0,V_0 for this problem must exist (but are not necessarily unique) because the sets of unitary matrices in M_n and M_m are compact and the Cartesian product of compact sets is compact. The maximizing matrices U_0,V_0 have the property that

$$\operatorname{Re}\operatorname{tr}(AV_0^*B^*)U_0^*\geq\operatorname{Re}\operatorname{tr}(AV_0^*B^*)U$$

for any unitary matrix $U\in M_m$, so by (7.4.9) we know that $AV_0^*B^*U_0^*$ is positive semidefinite. By the same argument,

$$\operatorname{Re} \operatorname{tr} AV_0^* B^* U_0^* = \operatorname{Re} \operatorname{tr} (B^* U_0^* A) V_0^* \geq \operatorname{Re} \operatorname{tr} B^* U_0^* AV$$

for any unitary matrix $V \in M_n$, so by (7.4.9) again we know that $B^* U_0^* A V_0^*$ is positive semidefinite. Thus, the two matrices $AV_0^* \in M_{m,n}$ and $B^* U_0^* \in M_{n,m}$ satisfy the hypotheses of Theorem (7.4.10) and hence if $q = \min\{m, n\}$ we have

$$\max\{\operatorname{Re} \operatorname{tr} AV^*B^*U^* : U \in M_m \text{ and } V \in M_n \text{ are unitary}\}$$

$$= \operatorname{Re} \operatorname{tr} AV_0^* B^* U_0^* = \sum_{i=1}^{q} \sigma_i(AV_0^*) \sigma_{\tau(i)}(B^* U_0^*) = \sum_{i=1}^{q} \sigma_i(A) \sigma_{\tau(i)}(B)$$

for some permutation τ of the integers $1, \ldots, q$ since the singular values are unitarily invariant. There is no loss of generality if we write the singular values $\sigma_1(A), \ldots, \sigma_q(A)$ and $\sigma_1(B), \ldots, \sigma_q(B)$ in decreasing order. If the permutation τ is not the identity, there are indices i_1, i_2 with $1 \leq i_1 < i_2 \leq q$ for which $\sigma_{\tau(i_1)}(B) \leq \sigma_{\tau(i_2)}(B)$, and one checks easily that the sum

$$\sum_{i=1}^{q} \sigma_i(A) \sigma_{\tau(i)}(B)$$

is not decreased if the permutation is altered to exchange the positions of these two singular values. In fact, the difference between the new and old values of the sum is

$$[\sigma_{i_1}(A) - \sigma_{i_2}(A)][\sigma_{\tau(i_2)}(B) - \sigma_{\tau(i_1)}(B)] \geq 0$$

Thus, the maximum value of the sum is achieved for the identity permutation τ, and we can conclude that

$$\max\{\operatorname{Re} \operatorname{tr} AV^*B^*U^* : U \in M_m, V \in M_n \text{ are unitary}\} = \sum_{i=1}^{q} \sigma_i(A) \sigma_i(B) \tag{7.4.14}$$

where the singular values of A and B are both arranged in decreasing order.

Using this result in our original minimization problem, we find for $A, B \in M_{m,n}$ and $q = \min\{m, n\}$ that

$$\min\{\|A - UBV\|_2 : U \in M_m \text{ and } V \in M_n \text{ are unitary}\}$$

$$= [\|A\|_2^2 - 2 \sum_{i=1}^{q} \sigma_i(A) \sigma_i(B) + \|B\|_2^2]^{1/2}$$

$$= \left[\sum_{i=1}^{q} \sigma_i^2(A) - 2 \sum_{i=1}^{q} \sigma_i(A) \sigma_i(B) + \sum_{i=1}^{q} \sigma_i^2(B) \right]^{1/2} \tag{7.4.15}$$

$$= \left[\sum_{i=1}^{q} [\sigma_i(A) - \sigma_i(B)]^2 \right]^{1/2}$$

In particular, A is a "two-sided rotation" of B if and only if A and B have the same set of singular values. □

Exercise. What does (7.4.15) say if $B = I$? Compare with the result at the end of example (7.4.8). What does (7.4.15) say if B is diagonal and has rank k? Compare with the comments in Example (7.4.1).

7.4.16 **Example.** As another example of the use of singular values, we consider the question of characterizing unitarily invariant norms of matrices, which were introduced in Section (5.6).

Definition: A vector norm $\|\cdot\|$ on $M_{m,n}$ is said to be *unitarily invariant* if

$$\|UAV\| = \|A\|$$

for all $A \in M_{m,n}$ and for all unitary matrices $U \in M_m$, $V \in M_n$.

If $A \in M_{m,n}$ is a given matrix and if $A = V\Sigma W^*$ is a singular value decomposition of A, then $\|A\| = \|V\Sigma W^*\| = \|\Sigma\|$ for any unitarily invariant norm $\|\cdot\|$. Thus, a unitarily invariant norm of a matrix of a given size depends only on the set of singular values of the matrix.

Two familiar examples of unitarily invariant norms are the Frobenius (Euclidean) norm and the spectral norm. If the singular values of $X = [x_{ij}] \in M_{m,n}$ are $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_q \geq 0$ $(q = \min\{m, n\})$, then

$$\|X\|_2 = \left(\sum_{j=1}^{n} \sum_{i=1}^{m} |x_{ij}|^2 \right)^{1/2} = \left(\sum_{i=1}^{q} \sigma_i^2 \right)^{1/2}$$

and

$$\|\!|X|\!\|_2 = \max_{y \neq 0} \frac{\|Xy\|_2}{\|y\|_2} = [\rho(X^*X)]^{1/2} = \sigma_1 = \max\{\sigma_1, \ldots, \sigma_q\}$$

For a general unitarily invariant norm $\|\cdot\|$ on $M_{m,n}$, the nature of its dependence on the singular values of its argument is easily determined. For convenience, assume that $m \leq n$, let $A = \operatorname{diag}(x_1, x_2, \ldots, x_m) \in M_m$, and define the partitioned matrix

$$X = [A \mid 0], \qquad A \in M_m, \qquad 0 \in M_{m, n-m}$$

Since $XX^* = \operatorname{diag}(|x_1|^2, |x_2|^2, \ldots, |x_m|^2)$, the set of singular values of X is $\{\sigma_i\} = \{|x_i|\}$. If we define the function $g : \mathbf{C}^m \to \mathbf{R}^+$ by

$$g(x) = g([x_1, \ldots, x_m]^T) \equiv \|X\|$$

then the function $g(\cdot)$ inherits certain properties from the norm $\|\cdot\|$:

(7.4.17) $g(x) \geq 0$ for all $x \in \mathbf{C}^m$ since $\|X\| \geq 0$ for all $X \in M_{m,n}$

(7.4.18) $g(x) = 0$ if and only if $x = 0$ since $\|X\| = 0$ if and only if $X = 0$.

(7.4.19) $g(\alpha x)=|\alpha|g(x)$ for all $x\in\mathbf{C}^m$ and all $\alpha\in\mathbf{C}$ since $\|\alpha X\|=|\alpha|\,\|X\|$ for all $\alpha\in\mathbf{C}$ and all $X\in M_{m,n}$

(7.4.20) $g(x+y)\le g(x)+g(y)$ for all $x,y\in\mathbf{C}^m$ since $\|X+Y\|\le\|X\|+\|Y\|$ for all $X,Y\in M_{m,n}$

These four properties say that $g(\bullet)$ must be a vector norm on $\mathbf{C}^m$, but $g(\bullet)$ has two additional properties:

(7.4.21) $g(\bullet)$ is an *absolute norm* on $\mathbf{C}^m$, as defined in (5.5.9); that is, if $x=[x_i]\in\mathbf{C}^m$ and if $y=[y_i]\equiv[|x_i|]\in\mathbf{C}^m$, then $g(x)=g(y)$. This is because $g(x)=\|X\|$ depends only on the singular values of X, which are $\sigma_i=|x_i|$.

(7.4.22) If $P\in M_m$ is a permutation matrix, then $g(Px)=g(x)$ for all $x\in\mathbf{C}^m$ because the set of singular values of $X=[A\,|\,0]$ is the same as that of $[PA\,|\,0]$ since $(PA)^*(PA)=A^*P^TPA=A^*A$. The function $g(x)$ is a function of the *set* of absolute values of the components of x without regard to their ordering.

Exercise. Compute an explicit singular value decomposition for the matrix $X=[A\,|\,0]\in M_{m,n}$, $A=\text{diag}(x_1,\ldots,x_m)$, which we have been discussing.

Exercise. If $m\ge n$, let $X=[A\,|\,0]^T$ with $A=\text{diag}(x_1,\ldots,x_n)\in M_n$ and define $g(x)=\|X\|$ for $x\in\mathbf{C}^n$. If $\|\bullet\|$ is a unitarily invariant norm on $M_{m,n}$ show that $g(\bullet)$ is an absolute vector norm on $\mathbf{C}^n$ such that $g(Px)=g(x)$ for all $x\in\mathbf{C}^n$ and every permutation matrix $P\in M_n$.

Exercise. Show directly that the vector norms $g(\bullet)$ derived from the Frobenius and spectral norms satisfy the six properties (7.4.17)–(7.4.22) above.

7.4.23 **Definition.** A function $g(\bullet):\mathbf{C}^q\to\mathbf{R}^+$ is said to be a *symmetric gauge function* if and only if it satisfies the six properties (7.4.17)–(7.4.22) above, that is, if and only if it is an absolute vector norm that is a permutation invariant function of the entries of its argument.

We have seen that every unitarily invariant norm on $M_{m,n}$ generates a symmetric gauge function; it is more interesting that the converse is true as well. The following theorem says that a function $N(\bullet)$ on $M_{m,n}$ is a unitarily invariant norm if and only if $N(A)$ is a symmetric gauge function of the singular values of A.

7.4.24 **Theorem.** Let $\|\bullet\|$ be a unitarily invariant norm on $M_{m,n}$, let $q=\min\{m,n\}$, let $x=[x_i]\in\mathbf{C}^q$, let $X_1=\text{diag}(x_1,\ldots,x_q)$, and let $X\equiv[X_1\,|\,0]\in M_{m,n}$ if $m\le n$ or $X\equiv[X_1\,|\,0]^T\in M_{m,n}$ if $m\ge n$. Let $g:\mathbf{C}^q\to\mathbf{R}^+$ be defined by $g(x)\equiv\|X\|$. Then $g(\bullet)$ is a symmetric gauge function. Conversely, if $g:\mathbf{C}^q\to\mathbf{R}^+$ is a given symmetric gauge function, and if $\|\bullet\|:M_{m,n}\to\mathbf{R}^+$ is defined by $\|A\|\equiv g([\sigma_1,\ldots,\sigma_q]^T)$, where $\sigma_1,\ldots,\sigma_q$ are the singular values of A, then $\|\bullet\|$ is a unitarily invariant norm on $M_{m,n}$.

Proof: The forward assertions have already been proved. For the converse, observe that $\|\bullet\|$ is a well-defined function on $M_{m,n}$ because $g(\bullet)$ is a permutation-invariant function of the components of its argument. Because the set of singular values of a matrix is a unitary invariant, we also have $\|UAV\|=\|A\|$ for all unitary $U\in M_m$ and $V\in M_n$. Because $g(\bullet)$ is a vector norm, we have $\|A\|\ge 0$ for all $A\in M_{m,n}$, and $\|A\|=0$ if and only if $g([\sigma_1,\ldots,\sigma_q]^T)=0$, and this can happen if and only if all $\sigma_i=0$ since $g(\bullet)$ is positive (7.4.18). But the zero matrix is the only matrix whose singular values are all zero, so the function $\|\bullet\|$ is positive (5.1.1(1a)). It is also homogeneous since $\sigma_i(cA)=|c|\sigma_i(A)$ and hence $\|cA\|=g([|c|\sigma_1,\ldots,|c|\sigma_q]^T)=|c|g([\sigma_1,\ldots,\sigma_q]^T)=|c|\,\|A\|$. What we have shown so far is that any function $\|\bullet\|$ generated in this way by a symmetric gauge function is a pre-norm on $M_{m,n}$ [see (5.4)]. It remains to be shown that $\|\bullet\|$ satisfies the triangle inequality, and we shall do so by showing that $\|\bullet\|$ is the dual of a pre-norm and hence [see the discussion following (5.4.12)] is actually a norm.

Consider the dual $g^D(\bullet)$ of the norm $g(\bullet)$ on $\mathbf{C}^q$:

$$g^D(y)\equiv\max_{g(x)=1}\text{Re}\,y^*x \tag{7.4.25}$$

The function $g^D(\bullet)$ is always a norm because $g(\bullet)$ is a (pre-)norm, but it is also a symmetric gauge function because

(7.4.21′) If $E=\text{diag}(e^{i\theta_1},\ldots,e^{i\theta_q})$ with all $\theta_i\in\mathbf{R}$, then

$$g^D(Ey)=\max_{g(x)=1}\text{Re}(Ey)^*x=\max_{g(x)=1}\text{Re}\,y^*(\bar{E}x)=\max_{g(Ex)=1}\text{Re}\,y^*x$$

$$=\max_{g(x)=1}\text{Re}\,y^*x=g^D(y)$$

because $g(\bullet)$ satisfies (7.4.21). Thus, $g^D(\bullet)$ also satisfies (7.4.21).

(7.4.22′) The same argument shows that if $P\in M_q$ is a permutation matrix, then

$$g^D(Py)=\max_{g(x)=1}\text{Re}(Py)^*x=\max_{g(x)=1}\text{Re}\,y^*P^Tx=\max_{g(Px)=1}\text{Re}\,y^*x$$

$$=\max_{g(x)=1}\text{Re y*x½g}^D(y)$$

because $g(\bullet)$ satisfies (7.4.22).

Thus, we can define the function $\|\cdot\|^D$ on M_n associated with the symmetric gauge function $g^D(\cdot)$:

$$\|A\|^D \equiv g^D([\sigma_1, \ldots, \sigma_q]^T)$$

where $\sigma_1, \ldots, \sigma_q$ are the singular values of A. [There is a conscious abuse of notation here: $\|\cdot\|^D$ usually denotes the dual of the norm $\|\cdot\|$; although we do not yet know that $\|\cdot\|$ is a norm, we shall show that this is the case and that $\|\cdot\|^D$, as defined in terms of a symmetric gauge function $g^D(\cdot)$, is its dual.] We have already shown that this function $\|\cdot\|^D$ is a pre-norm on M_q, since it is defined in terms of a symmetric gauge function $g^D(\cdot)$.

Now compute the dual of $\|\cdot\|^D$, which is guaranteed to be a norm on $M_{m,n}$ by (5.4.12). Observe that a matrix $B \in M_{m,n}$ satisfies $\|B\|^D = 1$ if and only if a singular value decomposition of B is $B = V\Sigma W^*$ with unitary matrices $V \in M_m$ and $W \in M_n$, $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_q)$, and $g^D([\sigma_1, \ldots, \sigma_q]^T) = 1$. For each given matrix $A \in M_{m,n}$ we have

$$\begin{aligned}(\|A\|^D)^D &\equiv \max_{\|B\|^D=1} \operatorname{Re}[A, B] = \max_{\|B\|^D=1} \operatorname{Re} \operatorname{tr} AB^* \\ &= \max\{\operatorname{Re} \operatorname{tr} A(V\Sigma W^*)^* : V \in M_m \text{ and } W \in M_n \text{ are unitary}, \\ &\qquad \Sigma = \operatorname{diag}(s_1, \ldots, s_q), \text{ and} \\ &\qquad g^D([s_1, \ldots, s_q]^T) = 1\}\end{aligned}$$

For each diagonal matrix Σ satisfying the constraint, we can use (7.4.14) to evaluate the maximum value that can be achieved over all choices of unitary V, W:

$$(\|A\|^D)^D = \max\left\{\sum_{i=1}^{q} \sigma_i(A)\,|s_i| : g^D([s_1, \ldots, s_q]^T) = 1\right\}$$

But since all $\sigma_i(A) \geq 0$ it is apparent from Definition (5.4.12) that this maximum is exactly the dual norm of $g^D(\cdot)$ evaluated at the point $[\sigma_1(A), \ldots, \sigma_q(A)]^T$. The duality theorem (5.5.14), however, guarantees that the dual of the dual of a norm is the original norm, so

$$(\|A\|^D)^D = (g^D)^D([\sigma_1(A), \ldots, \sigma_q(A)]^T) = g([\sigma_1(A), \ldots, \sigma_q(A)]^T) \equiv \|A\|$$

Thus, for all $A \in M_{m,n}$, $\|A\| = (\|A\|^D)^D$, which guarantees that $\|\cdot\|$ is actually a norm, and hence it satisfies the triangle inequality. This conclusion also justifies our abuse of notation, since $(\|A\|)^D = ((\|A\|^D)^D)^D = \|A\|^D$ by the duality theorem. Thus, $\|\cdot\|^D$, defined via the symmetric gauge function $g^D(\cdot)$, is actually the same as the dual of the norm $\|\cdot\|$. □

An important and familiar example of a family of symmetric gauge functions on $\mathbf{C}^n$ is the family of l_p norms (5.2.4)

$$g([x_1, \ldots, x_n]^T) = \|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, \qquad 1 \leq p < \infty$$

When applied to the singular values of a matrix, as described in Theorem (7.4.24), the l_p norms generate unitarily invariant norms on $M_{m,n}$ known as *Schatten p norms*. The case $p = 2$ is the Frobenius (Euclidean) norm

$$\|A\|_2 = \left[\sum_i \sigma_i(A)^2\right]^{1/2}$$

the limiting case $p \to \infty$ is the spectral norm

$$|||A|||_2 = \max\{\sigma_i(A)\}$$

and the case $p = 1$ is the *trace norm*

$$|||A|||_{\mathrm{tr}} = \sum_i \sigma_i(A)$$

The trace norm arose naturally in Example (7.4.6) when we considered the problem of approximating a given square matrix by a scalar multiple of a unitary matrix.

Another family of symmetric gauge functions on $\mathbf{C}^n$, which also includes the trace norm and the spectral norm, is given in (7.4.44).

7.4.26 **Example.** Singular values play an important role in deriving an inequality of Wielandt that gives a geometric meaning to the condition number of a square nonsingular matrix with respect to the spectral norm.

Let $A \in M_n$ be a nonsingular matrix, let $B \equiv A^*A \in M_n$, and denote the singular values of A by $\sigma_1 \geq \cdots \geq \sigma_n > 0$. The eigenvalues of the positive definite matrix B (arranged in the conventional increasing order) are $0 < \sigma_n^2 \leq \sigma_{n-1}^2 \leq \cdots \leq \sigma_1^2$. Let $x, y \in \mathbf{C}^n$ be any pair of orthonormal vectors, define $C \equiv [x\, y]^* B [x\, y] \in M_2$, and denote the eigenvalues of C by $0 < \gamma_1 \leq \gamma_2$. The Poincaré separation theorem (4.3.16) with $r = 2$ says that

$$\lambda_k(B) = \sigma_{n-k+1}^2 \leq \lambda_k(C) = \gamma_k \leq \lambda_{n+k-2}(B) = \sigma_{3-k}^2, \qquad k = 1, 2$$

or

$$\sigma_n^2 \leq \gamma_1 \leq \sigma_2^2 \quad \text{and} \quad \sigma_{n-1}^2 \leq \gamma_2 \leq \sigma_1^2$$

For our purposes, the only interesting implication of these inequalities is

$$\sigma_n^2 \leq \gamma_1 \leq \gamma_2 \leq \sigma_1^2 \tag{7.4.27}$$

in which the first and last inequalities are equalities if x and y are orthonormal eigenvectors of B corresponding to eigenvalues that are squares of the largest and smallest singular values of A, respectively.

Compute

$$1-\frac{|x^*By|^2}{(x^*Bx)(y^*By)}=4\frac{(x^*Bx)(y^*By)-|x^*By|^2}{(x^*Bx+y^*By)^2-(x^*Bx-y^*By)^2}$$

$$=\frac{4\det C}{(\operatorname{tr} C)^2-(x^*Bx-y^*By)^2} \tag{7.4.28}$$

$$=\frac{4\gamma_1\gamma_2}{(\gamma_1+\gamma_2)^2-(x^*Bx-y^*By)^2}\geq\frac{4\gamma_1\gamma_2}{(\gamma_1+\gamma_2)^2}$$

with equality if and only if $x, y \in \mathbf{C}^n$ are orthonormal and $x^*Bx = y^*By$. We transform this inequality into the equivalent inequality

$$\frac{|x^*By|^2}{(x^*Bx)(y^*By)}\leq 1-\frac{4\gamma_1\gamma_2}{(\gamma_1+\gamma_2)^2}=\left(\frac{\gamma_1-\gamma_2}{\gamma_1+\gamma_2}\right)^2=\left(\frac{\gamma_2/\gamma_1-1}{\gamma_2/\gamma_1+1}\right)^2 \tag{7.4.29}$$

The upper bound in (7.4.29) is a monotonically increasing function of the ratio γ_2/γ_1 [as may be shown easily by observing that the derivative of the function $f(t)=(t-1)/(t+1)$ is positive for $t>0$]. By (7.4.27), this ratio has the upper bound σ_1^2/σ_n^2, and hence

$$\frac{|x^*By|^2}{(x^*Bx)(y^*By)}\leq\left(\frac{\sigma_1^2/\sigma_n^2-1}{\sigma_1^2/\sigma_n^2+1}\right)^2=\left(\frac{\kappa^2-1}{\kappa^2+1}\right)^2 \tag{7.4.30}$$

where we have introduced the positive parameter $\kappa=\kappa(A)=\sigma_1/\sigma_n=\|A\|_2\|A^{-1}\|_2\geq 1$, which is the *condition number* of A with respect to the spectral norm. If $u_1, u_n \in \mathbf{C}^n$ are orthonormal eigenvectors of B corresponding to the eigenvalues σ_1^2 and σ_n^2, respectively, and if $x=(u_1+u_n)/\sqrt{2}$, $y=(u_1-u_n)/\sqrt{2}$, then $\{x, y\}$ is an orthonormal set, $x^*Bx=y^*By=(\sigma_1^2+\sigma_n^2)/2$, and $x^*By=(\sigma_1^2-\sigma_n^2)/2$, so equality is attained in (7.4.30) in this case.

Define the angle θ in the first quadrant by $\cot(\theta/2)=\kappa$, so that

$$\frac{\kappa^2-1}{\kappa^2+1}=\frac{\cot^2(\theta/2)-1}{\cot^2(\theta/2)+1}=\frac{\cos^2(\theta/2)-\sin^2(\theta/2)}{\cos^2(\theta/2)+\sin^2(\theta/2)}=\cos\theta$$

and (7.4.30) can be written in the form

$$\frac{|x^*By|^2}{(x^*Bx)(y^*By)}\leq\cos^2\theta \tag{7.4.31}$$

If we now observe that the left-hand side of this inequality is homogeneous of degree 0 in both x and y, we can finally state Wielandt's inequality in two equivalent forms:

7.4.32 **Theorem.** Let $A \in M_n$ be a given nonsingular matrix with spectral condition number κ, and define the angle θ in the first quadrant by $\cot(\theta/2)=\kappa$. Then

$$|\langle Ax, Ay\rangle|\leq\cos\theta\|Ax\|_2\|Ay\|_2 \tag{7.4.33}$$

for every pair of orthogonal vectors $x, y \in \mathbf{C}^n$, where $\langle u, v\rangle \equiv v^*u$ denotes the Euclidean inner product and $\|u\|_2=(u^*u)^{1/2}$ denotes the Euclidean norm. Moreover, there exists an orthonormal pair of vectors $x, y \in \mathbf{C}^n$ for which equality holds in (7.4.33).

7.4.34 **Theorem.** Let $B \in M_n$ be a given positive definite matrix with eigenvalues $0<\lambda_1\leq\lambda_2\leq\cdots\leq\lambda_n$. Then

$$|x^*By|^2\leq\left(\frac{\lambda_n-\lambda_1}{\lambda_n+\lambda_1}\right)^2(x^*Bx)(y^*By) \tag{7.4.35}$$

for every pair of orthogonal vectors $x, y \in \mathbf{C}^n$. Moreover, there exists an orthonormal pair of vectors $x, y \in \mathbf{C}^n$ for which equality holds in (7.4.35).

Proofs: The inequality (7.4.33) follows from (7.4.31) by substituting $B=A^*A$. The inequality (7.4.35) follows from (7.4.30) by substituting $\sigma_i^2=\lambda_{n-i+1}$ and recognizing that every positive definite matrix B is of the form $B=A^*A$ for some nonsingular $A \in M_n$; one may take $A=B^{1/2}$. We have already observed that equality can be attained in (7.4.30) by an orthonormal pair. □

Exercise. Show that (7.4.35) is an improvement on the general Cauchy–Schwarz inequality, which is $|x^*By|=|\langle Cy, Cx\rangle|\leq\|Cx\|_2\|Cy\|_2$, where $C=B^{1/2}$. However, the Cauchy-Schwarz inequality applies to all pairs x, y, whereas (7.4.35) applies only to orthogonal pairs. What happens if $\lambda_1=\lambda_n$?

The form (7.4.33) of Wielandt's inequality leads immediately to a useful geometrical interpretation of the spectral condition number. If $x, y \in \mathbf{C}^n$ is an arbitrary orthonormal pair, the left-hand side of the inequality

$$\frac{|\langle Ax, Ay\rangle|}{\|Ax\|_2\|Ay\|_2}\leq\cos\theta \tag{7.4.36}$$

is the ordinary cosine of the smaller Euclidean angle between the nonzero vectors Ax and Ay. This bound says that the smaller angle between Ax and Ay is at most $\theta=\theta(A)$, where $\theta(A)$ is defined by $\cot[\theta(A)/2]=\kappa(A)$. Since equality is possible in this bound, we have established the geometrical interpretation of $\theta(A)$ as the minimum angle between Ax and Ay as x and y range over all possible orthonormal pairs of vectors. This point was discussed in Sections (5.8) and (6.3).

A well-known inequality of Kantorovich follows easily from Wielandt's inequality. For any $x \in \mathbf{C}^n$, define

$$y \equiv \|x\|_2^2(B^{-1}x) - (x^*B^{-1}x)x \tag{7.4.37}$$

and notice that $x^*y = 0$. Compute

$$By = \|x\|_2^2 x - (x^*B^{-1}x)Bx$$

$$x^*By = \|x\|_2^4 - (x^*B^{-1}x)(x^*Bx)$$

$$y^*By = -(x^*B^{-1}x)(y^*Bx)$$

Since B, and hence B^{-1} also, is positive definite, we must have $y^*By \geq 0$ and hence $y^*Bx = x^*By \leq 0$. Write the inequality (7.4.31) in the form

$$|x^*By|^2 \leq \cos^2\theta (x^*Bx)(y^*By)$$

and substitute the values for the particular choice (7.4.37) of the pair x, y to obtain

$$|x^*By|^2 \leq (\cos^2\theta)(x^*Bx)(x^*B^{-1}x)(-x^*By)$$

In either of the two possible cases $x^*By < 0$ or $x^*By = 0$, this implies that

$$-x^*By = -[\|x\|_2^4 - (x^*B^{-1}x)(x^*Bx)] \leq (\cos^2\theta)(x^*Bx)(x^*B^{-1}x)$$

or

$$(\sin^2\theta)(x^*Bx)(x^*B^{-1}x) \leq \|x\|_2^4 \tag{7.4.38}$$

for any $x \in \mathbf{C}^n$. Notice that (7.4.38) is an equality if $x = u_1 + u_n$ is the sum of orthonormal eigenvectors of B corresponding to its smallest and largest eigenvalues. This leads to two equivalent forms of Kantorovich's inequality corresponding to the two forms of Wielandt's inequality.

7.4.39 **Theorem.** Let $A \in M_n$ be a given nonsingular matrix with spectral condition number κ and define the angle θ in the first quadrant by $\cot(\theta/2) = \kappa$. Then

$$\|x\|_2^2 \geq \sin\theta \|Ax\|_2 \|(A^*)^{-1}x\|_2 \tag{7.4.40}$$

for all $x \in \mathbf{C}^n$. Moreover, there is a unit vector x for which (7.4.40) is an equality.

7.4.41 **Theorem.** Let $B \in M_n$ be a given positive definite matrix with eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. Then

$$\|x\|_2^4 \geq \frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2}(x^*Bx)(x^*B^{-1}x) \tag{7.4.42}$$

for all $x \in \mathbf{C}^n$. Moreover, there is a unit vector x for which (7.4.42) is an equality.

Proofs: These two results follow from (7.4.38) by substituting $B = A^*A$ into (7.4.38) and by recognizing that

$$\sin^2\theta = 1 - \cos^2\theta = 1 - \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2 = \frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2}$$

The fact that equality is possible in both cases follows from the case of equality for (7.4.38). □

7.4.43 **Example.** It is sometimes possible to prove norm inequalities for matrices that hold for all unitarily invariant norms. The key to doing so lies in recognizing the fundamental role of the particular symmetric gauge functions $g_k([x_1, \ldots, x_n]^T)$ defined on $\mathbf{C}^n$ by

$$g_k(x) = \max\{|x_{i_1}| + \cdots + |x_{i_k}| : 1 \leq i_1 < i_2 < \cdots < i_k \leq n\}, \qquad k = 1, \ldots, n \tag{7.4.44}$$

When applied to the singular values of a matrix, as described in Theorem (7.4.24), this particular family of symmetric gauge functions generates a family of unitarily invariant norms on $M_{m,n}$ known as *Ky Fan k norms.* The case $k = 1$ is the spectral norm and the case $k = \min\{m, n\}$ is the trace norm.

7.4.45 **Theorem.** Let $x = [x_i]$, $y = [y_i] \in \mathbf{C}^n$ be given vectors. Then $g(x) \leq g(y)$ for all symmetric gauge functions $g(\cdot)$ on $\mathbf{C}^n$ if and only if $g_k(x) \leq g_k(y)$ for $k = 1, 2, \ldots, n$, where the $g_k(\cdot)$ are the particular symmetric gauge functions defined in (7.4.44).

Proof: Since each $g_k(\cdot)$ is a symmetric gauge function, the necessity of the condition is clear. To prove sufficiency, suppose $g_k(x) \leq g_k(y)$ for $k = 1, 2, \ldots, n$ and let $g(\cdot)$ be a given symmetric gauge function. Because a symmetric gauge function is a permutation-invariant function of the components of its argument (7.4.22), there is no loss of generality if we assume for convenience that the absolute values of the components of x and y are each arranged in nondecreasing order

$$|x_1| \leq |x_2| \leq \cdots \leq |x_n|, \qquad |y_1| \leq |y_2| \leq \cdots \leq |y_n|$$

The assumption that $g_k(x) \leq g_k(y)$ for all $k = 1, 2, \ldots, n$ is then equivalent to the set of n inequalities

$$\begin{aligned}
|x_n| &\le |y_n| \\
|x_{n-1}|+|x_n| &\le |y_{n-1}|+|y_n| \\
&\vdots \\
|x_2|+\cdots+|x_n| &\le |y_2|+\cdots+|y_n| \\
|x_1|+|x_2|+\cdots+|x_n| &\le |y_1|+|y_2|+\cdots+|y_n| \quad (*)
\end{aligned} \qquad (7.4.46)$$

The similarity between these inequalities and the defining inequalities (4.3.24) for majorization is not merely superficial.

If the last of these inequalities $(*)$ is not an equality, modify y by decreasing the absolute value of the component y_1 until either (a) inequality $(*)$ is an equality or (b) $|y_1|$ is decreased to zero. If (b) occurs before (a), repeat this process with the next component y_2, and so on until (a) occurs. The result will be to produce a modified vector $y'=[y_i']$ such that $|y_i'|\le|y_i|$ for $i=1,\ldots,n$, $g_k(x)\le g_k(y')$ for all $k=1,\ldots,n$, and equality holds in $(*)$. Because an absolute norm is also a monotone norm (5.5.10), we have $g(y')\le g(y)$. Thus, if we can show that $g(x)\le g(y)$ for any $x,y\in\mathbf{C}^n$ that satisfy the inequalities (7.4.46) and for which $(*)$ is an equality, we can conclude that $g(x)\le g(y)$ for any $x,y\in\mathbf{C}^n$ that satisfy (7.4.46) in general.

The assumption that (7.4.46) holds with equality in $(*)$ is exactly the assumption that the vector $-|x|=[-|x_i|]\in\mathbf{R}^n$ majorizes the vector $-|y|=[-|y_i|]\in\mathbf{R}^n$ (4.3.24), and in this event we know (4.3.33) that there is a doubly stochastic matrix $S\in M_n$ such that $-|x|=S(-|y|)$, or $|x|=S|y|$. Since every doubly stochastic matrix is a convex combination of finitely many permutation matrices (8.7.1), we can write $S=\alpha_1P_1+\cdots+\alpha_NP_N$, where $\alpha_i\ge0$, $\alpha_1+\cdots+\alpha_N=1$, and each $P_i\in M_n$ is a permutation matrix. In this event, we have

$$\begin{aligned}
g(x)&=g(|x|)\\
&=g(S|y|)=g\left(\sum_{i=1}^{N}\alpha_iP_i|y|\right)\le\sum_{i=1}^{n}g(\alpha_iP_i|y|)=\sum_{i=1}^{N}\alpha_i g(|y|)=g(|y|)\\
&=g(y)
\end{aligned}$$

because $g(\cdot)$ is an absolute vector norm that is a permutation-invariant function of the components of its argument. $\square$

The significance of the theorem is that in order to have $\|A\|\le\|B\|$ for *every* unitarily invariant norm $\|\cdot\|$ on $M_{m,n}$, it is necessary and sufficient that this inequality hold for the Ky Fan k norms, $k=1,2,\ldots,\min\{m,n\}$.

7.4.47 **Corollary.** Let $A,B\in M_{m,n}$ be given matrices with respective singular values $\sigma_1(A)\ge\cdots\ge\sigma_q(A)\ge0$ and $\sigma_1(B)\ge\cdots\ge\sigma_q(B)\ge0$, where $q=\min\{m,n\}$. In order that $\|A\|\le\|B\|$ for every unitarily invariant norm $\|\cdot\|$ on $M_{m,n}$ it is sufficient that

$$\sigma_i(A)\le\sigma_i(B)\quad\text{for all}\quad i=1,2,\ldots,q \qquad (7.4.48)$$

and it is necessary and sufficient that

$$\begin{aligned}
\sigma_1(A)&\le\sigma_1(B)\\
\sigma_1(A)+\sigma_2(A)&\le\sigma_1(B)+\sigma_2(B)\\
&\vdots\\
\sigma_1(A)+\sigma_2(A)+\cdots+\sigma_q(A)&\le\sigma_1(B)+\cdots+\sigma_q(B)
\end{aligned} \qquad (7.4.49)$$

Proof: The key observation required is that a unitarily invariant norm on $M_{m,n}$ is a symmetric gauge function of the singular values of its argument (7.4.24). The sufficiency of (7.4.48) requires only the fact that a symmetric gauge function is a monotone norm (5.5.10), while the more subtle assertion about the inequalities (7.4.49) is the content of the preceding theorem. $\square$

In order to apply Corollary (7.4.47) to prove norm inequalities, it is frequently useful to have the following restatement of the fact that the vector of eigenvalues of a sum of Hermitian matrices majorizes the sum of the respective vectors of ordered eigenvalues.

7.4.50 **Lemma.** Let $A,B\in M_n$ be Hermitian matrices with ordered eigenvalues $\lambda_1(A)\le\cdots\le\lambda_n(A)$ and $\lambda_1(B)\le\cdots\le\lambda_n(B)$, and let $\lambda_1(A-B)\le\cdots\le\lambda_n(A-B)$ denote the ordered eigenvalues of $A-B$. Then the vector

$$\lambda(A)-\lambda(B)=[\lambda_i(A)-\lambda_i(B)]$$

majorizes the vector $\lambda(A-B)=[\lambda_i(A-B)]$; that is,

$$\min\left\{\sum_{j=1}^{k}[\lambda_{i_j}(A)-\lambda_{i_j}(B)]:1\le i_1<i_2<\cdots<i_k\le n\right\}\ge\sum_{i=1}^{k}\lambda_i(A-B)$$

for $k=1,2,\ldots,n$, with equality for $k=n$.

Proof: Theorem (4.3.27) says that the vector $\lambda(A)=\lambda((A-B)+B)=[\lambda_i((A-B)+B)]$ of eigenvalues of $(A-B)+B=A$ majorizes the vector $\lambda(A-B)+\lambda(B)=[\lambda_i(A-B)+\lambda_i(B)]$, which is equivalent to having the vector $\lambda(A)-\lambda(B)$ majorize the vector $\lambda(A-B)$. $\square$

With the conditions in (7.4.47) and the preceding lemma as tools, it is often possible to generalize approximation theorems or inequalities for the Frobenius norm or the spectral norm to the whole class of unitarily invariant norms.

For example, (7.4.15) says that if $A, B \in M_{m,n}$ are given matrices with respective ordered singular values $\sigma_1(A) \geq \cdots \geq \sigma_q(A) \geq 0$ and $\sigma_1(B) \geq \cdots \geq \sigma_q(B) \geq 0$, with $q = \min\{m, n\}$, then

$$\|A-B\|_2 \geq \left(\sum_{i=1}^{q} [\sigma_i(A)-\sigma_i(B)]^2\right)^{1/2}$$

Another way to write this lower bound is

$$\|A-B\|_2 \geq \|\Sigma(A)-\Sigma(B)\|_2$$

where $A = V_1\Sigma(A)W_1^*$ and $B = V_2\Sigma(B)W_2^*$ are singular value decompositions in which the respective singular values are ordered from largest to smallest on the "diagonal" of $\Sigma(A)$ and $\Sigma(B)$. Another instance of this inequality, for the spectral norm, is (7.3.8(a)). It is in this form that (7.4.15) generalizes to all unitarily invariant norms.

7.4.51 **Theorem.** Let $A, B \in M_{m,n}$ be given matrices with singular value decompositions $A = V_1\Sigma(A)W_1^*$ and $B = V_2\Sigma(B)W_2^*$ with unitary $V_1, V_2 \in M_m$ and unitary $W_1, W_2 \in M_n$ and in which the "diagonal" elements of both $\Sigma(A)$ and $\Sigma(B)$ are arranged in decreasing order. Then $\|A-B\| \geq \|\Sigma(A)-\Sigma(B)\|$ for every unitarily invariant norm $\|\bullet\|$ on $M_{m,n}$.

Proof: Let $q = \min\{m, n\}$. Use (7.3.7) to identify the singular values of A

$$\sigma_1(A) \geq \cdots \geq \sigma_q(A) \geq 0$$

with the first q nonpositive eigenvalues of the Hermitian matrix

$$\tilde{A} = \begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} \in M_{m+n}$$

of which the $m+n$ ordered eigenvalues are

$$-\sigma_1(A) \leq -\sigma_2(A) \leq \cdots \leq -\sigma_q(A) \leq 0 = \cdots = 0 \leq \sigma_q(A) \leq \cdots \leq \sigma_1(A)$$

and similarly for $\tilde{B}$ and $\tilde{A}-\tilde{B}$. The differences of the ordered eigenvalues of $\tilde{A}$ and $\tilde{B}$ are $\pm[\sigma_1(A)-\sigma_1(B)], \ldots, \pm[\sigma_q(A)-\sigma_q(B)]$ together with 0 ($|m-n|$ times). Although it is not clear how to order this sequence in general, the q smallest elements in an ordering of this sequence are $\{-|\sigma_i(A)-\sigma_i(B)|\}$, and Lemma (7.4.50) applied to $\tilde{A}$, $\tilde{B}$, and $\tilde{A}-\tilde{B}$ assures us that

$$\sum_{i=1}^{k} -\sigma_i(A-B) \leq \min\left\{\sum_{j=1}^{k} -|\sigma_{i_j}(A)-\sigma_{i_j}(B)| : 1 \leq i_1 < \cdots < i_k \leq n\right\}$$

for $k = 1, \ldots, q$, which is equivalent to

$$\sum_{i=1}^{k} \sigma_i(A-B) \geq \max\left\{\sum_{j=1}^{k} |\sigma_{i_j}(A)-\sigma_{i_j}(B)| : 1 \leq i_i < \cdots < i_k \leq n\right\}$$

$k = 1, \ldots, q$. Since $\{|\sigma_i(A)-\sigma_i(B)|\}$ is the set of singular values of $\Sigma(A)-\Sigma(B)$, Corollary (7.4.47) guarantees that $\|A-B\| \geq \|\Sigma(A)-\Sigma(B)\|$ for any unitarily invariant norm $\|\bullet\|$. □

7.4.52 **Example.** One consequence of Theorem (7.4.51) is a generalization of the problem of finding a best (in the sense of least squares) rank k approximation to a given matrix $A \in M_n$, considered in Example (7.4.1) for the Frobenius norm. If $\|\bullet\|$ is a unitarily invariant norm and if $B \in M_n$ has rank k, then $\sigma_1(B) \geq \cdots \geq \sigma_k(B) > 0 = \sigma_{k+1}(B) = \cdots = \sigma_n(B)$. Thus,

$$\begin{aligned}\|A-B\| &\geq \|\Sigma(A)-\Sigma(B)\| \\ &= \|\operatorname{diag}(\sigma_1(A)-\sigma_1(B), \ldots, \sigma_k(A)-\sigma_k(B), \sigma_{k+1}(A), \ldots, \sigma_n(A))\| \\ &\geq \|\operatorname{diag}(0, \ldots, 0, \sigma_{k+1}(A), \ldots, \sigma_n(A))\|\end{aligned}$$

where we have used the fact that a unitarily invariant norm on diagonal matrices is a monotone norm because it is a symmetric gauge function of the diagonal entries. Furthermore, equality is possible for $B = VEW^*$, where $A = V\Sigma(A)W^*$ is a singular value decomposition for A and $E = \operatorname{diag}[\sigma_1(A), \ldots, \sigma_k(A), 0, \ldots, 0]$.

Thus, for any $A \in M_n$ and any $B \in M_n$ of rank k, we have the bounds

$$\begin{aligned}\|A-B\| &\geq \|\operatorname{diag}(0, \ldots, 0, \sigma_{k+1}(A), \ldots, \sigma_n(A))\| \\ &\geq \sigma_n(A)\|\operatorname{diag}(0, \ldots, 0, 1, \ldots, 1)\|\end{aligned}$$

for any unitarily invariant norm (there are k zero terms on the diagonal of the last expression) in which the first inequality, but not generally the second, is sharp. The second inequality (which follows solely from monotonicity of symmetric gauge functions if A is nonsingular and is trivial if A is singular) has the advantage that its dependence on the norm is a function of k only, and not of A. In particular, this says that for any nonsingular matrix $A \in M_n$ and any unitarily invariant norm $\|\bullet\|$, we have the sharp bound

$$\|A-B\| \geq \sigma_n(A)\|\operatorname{diag}(0, \ldots, 0, 1)\| \tag{7.4.53}$$

for the distance between A and any singular matrix B; that is, the minimum distance from A to the closed set of singular matrices (with respect to the unitarily invariant norm $\|\cdot\|$) is $\sigma_n(A)\|\text{diag}(0, \ldots, 0, 1)\|$.

7.4.54 **Example.** Properties of symmetric gauge functions can be exploited to give a simple characterization of the unitarily invariant norms on M_n that are matrix norms. If $\|\cdot\|$ is a unitarily invariant matrix norm on M_n, then we know from Corollary (5.6.35) that $\|A\| \geq \sigma_1(A)$ for all $A \in M_n$. Using Theorem (5.6.9) and the fact that every unitarily invariant norm on M_n is self-adjoint (see Problem 2), we can also prove this directly by observing that $[\sigma_1(A)]^2 = \rho(A^*A) \leq \|A^*A\| \leq \|A^*\|\,\|A\| = \|A\|^2$. On the other hand, let $\|\cdot\|$ be a unitarily invariant norm on M_n such that $\|A\| \geq \sigma_1(A)$ for all $A \in M_n$, and let g be the symmetric gauge function on $\mathbf{C}^n$ generated by $\|\cdot\|$. The multiplicative analog for singular values of Weyl's inequalities given in Problem 18 of Section (7.3) and the fact that g is a monotone norm imply that

$$\begin{aligned}\|AB\| &= g(\sigma_1(AB), \sigma_2(AB), \ldots, \sigma_n(AB))\\ &\leq g(\sigma_1(A)\sigma_1(B), \sigma_1(A)\sigma_2(B), \ldots, \sigma_1(A)\sigma_n(B))\\ &= \sigma_1(A)\, g(\sigma_1(B), \sigma_2(B), \ldots, \sigma_n(B))\\ &= \sigma_1(A)\|B\| \leq \|A\|\,\|B\|\end{aligned}$$

Thus, *a unitarily invariant norm* $\|\cdot\|$ *on* M_n *is a matrix norm if and only if* $\|A\| \geq \sigma_1(A) = \|A\|_2$ *for all* $A \in M_n$. In particular, all the Ky Fan k norms for $k = 1, 2, \ldots, n$ and all the Schatten p-norms for $p \geq 1$ [generated by the symmetric gauge functions in (7.4.4) and (5.2.4), respectively] are matrix norms. Another consequence of this characterization is that *the set of unitarily invariant matrix norms on* M_n *is convex.* The set of *all* matrix norms on M_n is not convex [see Problem 9 of Section (5.6)].

Problems

1. Let $A \in M_{m,n}$ have rank $k > 0$. Suppose it is desired to find a matrix $A_1 \in M_{m,n}$ that has rank $k_1 < k$ and most closely approximates A in the Frobenius norm. Show that this can be done as follows: Let $A = V\Sigma W^*$ be a singular value decomposition of A. Let Σ_1 be the same as Σ except that only $\sigma_1, \ldots, \sigma_{k_1}$ are used; the remaining $n - k_1$ "diagonal" entries of Σ_1 are zero. Then $A_1 \equiv V\Sigma_1 W^*$ has the required property. *Hint:* Use (7.4.15). Notice that (7.4.52) shows that the given approximation is "best" not only for the Frobenius norm, but for all unitarily invariant norms as well.

2. A norm $\|\cdot\|$ on M_n is said to be *self-adjoint* if $\|A\| = \|A^*\|$ for every $A \in M_n$. Use Theorem (7.4.24) to show that every unitarily invariant norm on M_n is self-adjoint. Given an example of a self-adjoint norm that is not unitarily invariant.

3. Use Theorem (7.4.10) and the methods of Example (7.4.6) to determine a best least-squares approximation to a given matrix $A \in M_{m,n}$ (with $m \leq n$) by a scalar multiple of a matrix $Y \in M_{m,n}$ with orthonormal rows. *Hint:* Show that such a matrix Y must have the form $Y = VDW$, where $V \in M_m$ and $W \in M_n$ are unitary, $D = [I \,|\, 0] \in M_{m,n}$, $I \in M_m$, and $0 \in M_{m,n-m}$. The problem of minimizing $\|A - cY\|_2^2$ is the same as the problem of minimizing $\|A\|_2^2 - (\text{Re tr}\, AY^*)^2/m$. If $A = V_1\Sigma W_1^*$ is a singular value decomposition of A, show that this minimization problem requires finding

$$\max \text{Re tr}\{\Sigma W D^* V: W \in M_n \text{ and } V \in M_m \text{ are unitary}\}$$

and use Theorem (7.4.10) to solve this problem as in Example (7.4.13). Show that the value of the error in this case has the same form as in Example (7.4.6).

4. Consider diagonal matrices $A, B \in M_n$ to show that all possible permutations τ can arise in (7.4.11).

5. Consider the function $u(A)$ defined in (7.4.7). Show that

$$u(A) \leq \sqrt{n}\|A\|_2 \quad \text{for all} \quad A \in M_n$$

and that this bound is sharp. Use the definition to show directly that $u(A)$ is a vector norm on M_n, and explain why $u(A)$ is actually a matrix norm on M_n. *Hint:* See Example (7.4.54).

6. Show that if $A \in M_n$ is nonsingular and if $\kappa(A) = \|A\|_2\|A^{-1}\|_2$ is the condition number of A with respect to the spectral norm, then $\kappa(A) = \sigma_1/\sigma_n$, the ratio of the largest and smallest singular values. How does this compare with the estimate $\kappa(A) \geq |\lambda_1/\lambda_n|$?

7. Show that the constant in Kantorovich's inequality (7.4.42) is the square of the ratio of the geometric mean of λ_1 and λ_n to the arithmetic mean of λ_1 and λ_n.

8. Let $A \in M_n$ be nonsingular and Hermitian. Use Kantorovich's inequality (7.4.40) to show that

$$\max_{x \neq 0} \frac{\|Ax\|_2\|A^{-1}x\|_2}{\|x\|_2^2} = \frac{\sigma_1^2 + \sigma_n^2}{2\sigma_1\sigma_n} = \frac{1}{2}\left(\frac{\sigma_1}{\sigma_n} + \frac{\sigma_n}{\sigma_1}\right)$$

where $\sigma_1 \geq \cdots \geq \sigma_n > 0$ are the singular values of A. Show that σ_1 and σ_n are the absolute values, respectively, of the eigenvalues of A of largest and smallest absolute value, and that

$$\frac{1}{2}\left(\frac{\sigma_1}{\sigma_n}+\frac{\sigma_n}{\sigma_1}\right)=\frac{1}{2}(\kappa+\kappa^{-1})$$

where κ is the spectral condition number of A. Exhibit a vector x for which the maximum is achieved. Use the definition of the spectral condition number and its relation to the above-defined maximum to explain why one must have

$$\frac{1}{2}\left(\frac{\sigma_1}{\sigma_n}+\frac{\sigma_n}{\sigma_1}\right)\leq\frac{\sigma_1}{\sigma_n}$$

Prove this inequality directly. *Hint:* Show that $f(x)=x-[x+(1/x)]/2$ is an increasing function for $x\geq 1$.

9. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be n given positive real numbers. Use Kantorovich's inequality (7.4.42) to prove that if $\alpha_1, \ldots, \alpha_n$ are nonnegative and sum to 1, then

$$\left(\sum_{i=1}^{n}\alpha_i\lambda_i\right)\left(\sum_{i=1}^{n}\frac{\alpha_i}{\lambda_i}\right)\leq\frac{(\lambda_{\max}+\lambda_{\min})^2}{4\lambda_{\max}\,\lambda_{\min}}$$

10. Prove the following generalization of Kantorovich's inequality (7.4.42) due to Greub and Rheinboldt: Let $B, C\in M_n$ be commuting positive definite matrices with eigenvalues $0<\lambda_1\leq\cdots\leq\lambda_n$ and $0<\mu_1\leq\cdots\leq\mu_n$, respectively. Then

$$(x^*BCx)^2\geq\frac{4\lambda_1\lambda_n\mu_1\mu_n}{(\lambda_1\mu_1+\lambda_n\mu_n)^2}(x^*B^2x)(x^*C^2x)$$

for all $x\in\mathbf{C}^n$. *Hint:* Since $B=U\Lambda U^*$ and $C=UMU^*$ for some unitary $U\in M_n$, write the asserted inequality first in terms of $y=U^*x$ and then in terms of $z=(\Lambda M)^{1/2}y$. Then apply (7.4.41) with $B=\Lambda M^{-1}$ to show that the asserted inequality holds (and is sharp) with a constant of the form

$$\frac{\lambda_1\lambda_n\mu_j\mu_k}{(\lambda_1\mu_j+\lambda_n\mu_k)^2}$$

for some choice of indices $1\leq j\neq k\leq n$. Show that the least constant of this form occurs for $j=1$ and $k=n$. This final generalized inequality might not be sharp, however.

11. In contrast to the Kantorovich inequality (7.4.42), show that

$$(x^*Bx)(x^*B^{-1}x)\geq\|x\|_2^4 \quad \text{for all} \quad x\in\mathbf{C}^n$$

if $B\in M_n$ is positive definite. More generally, show that

$$(x^*Bx)(y^*B^{-1}y)\geq(x^*y)^2 \quad \text{for all} \quad x, y\in\mathbf{C}^n$$

if $B\in M_n$ is positive definite, with equality for $x=B^{-1}y$. Conclude that

$$(x^*x)^2\leq(x^*Bx)(x^*B^{-1}x)\leq\frac{[(\lambda_1+\lambda_n)/2]^2}{\lambda_1\lambda_n}(x^*x)^2 \quad \text{for all} \quad x\in\mathbf{C}^n$$

Hint: Use the Cauchy-Schwartz inequality to show that

$$\left|\sum_{i=1}^{n}x_i\bar{y}_i\right|^2=\left|\sum_{i=1}^{n}(\sqrt{\lambda_i}\,x_i)\left(\frac{\bar{y}_i}{\sqrt{\lambda_i}}\right)\right|^2\leq\left(\sum_{i=1}^{n}\lambda_i|x_i|^2\right)\left(\sum_{i=1}^{n}\frac{|y_i|^2}{\lambda_i}\right)$$

if $\lambda_i>0$, and then write $B=U\Lambda U^*$.

12. Let $B\in M_n$ be positive definite, let $y\in\mathbf{C}^n$ be any nonzero vector, and define

$$f(B, y)\equiv\min\left\{\frac{x^*Bx}{(x^*y)^2}: x\in\mathbf{C}^n \text{ and } x^*y\neq 0\right\}$$

Show that $f(B, y)$ is well defined and use Problem 11 to show that $f(B, y)=1/y^*B^{-1}y$. Show that f has the superadditivity property

$$f(A+B, y)\geq f(A, y)+f(B, y)$$

for all nonzero $y\in\mathbf{C}^n$ and all positive definite $A, B\in M_n$. Now let $y=e_i$, the ith standard unit basis vector, and deduce Bergstrom's inequality

$$\frac{\det(A+B)}{\det(A_i+B_i)}\geq\frac{\det A}{\det A_i}+\frac{\det B}{\det B_i}, \qquad i=1, \ldots, n$$

valid for any positive definite matrices $A, B\in M_n$, where $A_i\in M_{n-1}$ denotes the principal submatrix of A obtained by deleting the ith row and column of A, and similarly for B_i. This approach to Bergstrom's inequality is an example of a widely useful technique known as *quasi-linearization:* express a nonlinear function of a quantity of interest as the constrained extremal value of another function that depends linearly (or perhaps only additively) on the quantity of interest. The crucial step in (7.4.24) (proving that the pre-norm on $M_{m,n}$ defined in terms of a symmetric gauge function of the singular values is actually a norm) was accomplished with the quasi-linearization in (5.4.12).

13. For any complex number z, the inequality $|z-\operatorname{Re} z|\leq|z-x|$ holds for all real numbers x. A plausible generalization of this to square matrices $A\in M_n$ is

$$\|A-\tfrac{1}{2}(A+A^*)\|\leq\|A-H\|$$

for all Hermitian matrices $H\in M_n$. Prove that this inequality holds for all unitarily invariant norms $\|\cdot\|$ and, more generally, for all self-adjoint norms. Conclude that the distance (with respect to $\|\cdot\|$) from a given matrix $A\in M_n$ to the closed set of Hermitian matrices in M_n is $\frac{1}{2}\|A-A^*\|$. *Hint:* $A-\frac{1}{2}(A+A^*)=\frac{1}{2}(A-H)+\frac{1}{2}(H-A^*)$, so $\|A-\frac{1}{2}(A+A^*)\|\leq\frac{1}{2}\|A-H\|+\frac{1}{2}\|H-A^*\|$.

14. For any complex number z, we have the inequality $|\mathrm{Re}\, z| \le |z|$. Demonstrate the trivial generalization $\|(A+A^*)/2\| \le \|A\|$ for all $A \in M_n$ and for all unitarily invariant (even self-adjoint) norms $\|\bullet\|$.

15. Let $A \in M_n$ be given, let $\lambda_1 \le \cdots \le \lambda_n$ be the ordered eigenvalues of $\frac{1}{2}(A+A^*)$, and let $\sigma_1 \ge \cdots \ge \sigma_n$ be the ordered singular values of A. Explain why the inequalities

$$\lambda_{n-k+1}(\tfrac{1}{2}[A+A^*]) \le \sigma_k(A), \qquad k=1,\ldots,n$$

may be thought of as a generalization of the inequality $\mathrm{Re}\, z \le |z|$ for complex numbers. This says that the kth largest singular value of A is greater than or equal to the kth largest eigenvalue of $\frac{1}{2}(A+A^*)$. *Hint:* If y is a Euclidean unit vector, then

$$\tfrac{1}{2}y^*(A+A^*)y = \mathrm{Re}\; y^*Ay \le \|Ay\|_2$$

Use the Courant–Fischer theorem (4.2.11) to express λ_{n-k+1}, and then use this inequality and (7.3.10) to get σ_k.

16. Let $A \in M_n$ be given, and let $\|\bullet\|$ be a unitarily invariant norm on M_n. Use (7.4.51) to show that $\|A-U\| \ge \|\Sigma(A)-I\|$ for any unitary $U \in M_n$ and that this inequality is sharp. Conclude that $\|\Sigma(A)-I\|$ is the distance (with respect to $\|\bullet\|$) from A to the compact set of unitary matrices in M_n.

17. Let $A \in M_n$ have a singular value decomposition $A = V\Sigma(A)W^*$ and let $\|\bullet\|$ be a unitarily invariant norm on M_n. Show that

$$\|\Sigma(A)-I\| \le \|A-U\| \le \|\Sigma(A)+I\|$$

for any unitary $U \in M_n$. *Hint:* Show that $\Sigma(U)=I$ in any singular value decomposition of any unitary matrix U, so that the lower bound follows immediately from (7.4.51). For the upper bound, use the singular value analog of Weyl's additive eigenvalue inequalities in Problem 16 in Section (7.3) to show that $\sigma_{i+j-1}(A+(-U)) \le \sigma_i(A)+\sigma_j(-U)$. Then use (7.4.48).

18. With the inequality in Example (7.4.53) for nonsingular A as a guide, find a sharp lower bound for $\|A-B\|$, where $A \in M_n$ is a given matrix of rank k_1, $B \in M_n$ is an arbitrary matrix of rank $k < k_1$, and $\|\bullet\|$ is a unitarily invariant norm.

Further Readings. The original version of Theorem (7.4.24) in the case $m=n$ is due to Von Neumann; see the paper cited in Section (5.4). The treatment given of the inequalities of Wielandt and Kantorovich is adapted from [Hou 64], which has many references to the original papers.

For generalizations and further references see A. Clausing, "Kantorovich-Type Inequalities," *Amer. Math. Monthly* 89 (1982), 314–320. The approach to Bergstrom's inequality in Problem 12 is taken from [BB], which has a long chapter (with copious references) devoted to inequalities arising from positive definite matrices; there is also a discussion, and many examples, of the method of quasi-linearization. For more information about inequalities valid for all unitarily invariant norms, see L. Mirsky, "Symmetric Gauge Functions and Unitarily Invariant Norms," *Quart. J. Math. Oxford* 11(2) (1960), 50–59 as well as K. Fan and A. J. Hoffman, "Some Metric Inequalities in the Space of Matrices," *Proc. Amer. Math. Soc.* 6 (1955), 111–116. As an example of how these results are applied in statistics, and for further references to the statistics literature, see C. R. Rao, "Matrix Approximations and Reduction of Dimensionality in Multivariate Statistical Analysis," *Multivariate Analysis-V*, Proceedings of the Fifth International Symposium on Multivariate Analysis, P. R. Krishnaiah, North-Holland, Amsterdam, 1980, pp. 1–22.

7.5 The Schur product theorem

A particularly simple (and seemingly naive) composition of matrices is component-wise multiplication.

7.5.1 **Definition.** If $A=[a_{ij}] \in M_{m,n}$ and $B=[b_{ij}] \in M_{m,n}$ are given, then the *Hadamard product* of A and B is the matrix $A \circ B \equiv [a_{ij}b_{ij}] \in M_{m,n}$.

The Hadamard product is often called the *Schur product*. Like matrix addition, Hadamard multiplication is commutative, and it is considerably simpler than the usual matrix multiplication.

The Hadamard product arises naturally from several different points of view. For example, if $f(\theta)$ and $g(\theta)$ are continuous periodic functions of period 2π and if

$$a_k = \int_0^{2\pi} e^{ik\theta} f(\theta)\, d\theta \quad \text{and} \quad b_k = \int_0^{2\pi} e^{ik\theta} g(\theta)\, d\theta$$

$k=0, \pm 1, \pm 2, \ldots$, then the convolution

$$h(\theta) \equiv \int_0^{2\pi} f(\theta - t) g(t)\, dt$$

has trigonometric moments

$$c_k = \int_0^{2\pi} e^{ik\theta} h(\theta)\, d\theta$$

that satisfy the identities $c_k = a_k b_k$, $k = 0, \pm 1, \pm 2, \ldots$. Thus, the Toeplitz matrix of trigonometric moments of $h(\theta)$ is the Hadamard product of the Toeplitz matrices of trigonometric moments of $f(\theta)$ and $g(\theta)$:

$$[c_{i-j}] = [a_{i-j}] \circ [b_{i-j}]$$

If $f(\theta)$ and $g(\theta)$ are both nonnegative real-valued functions, then the convolution $h(\theta)$ is also a nonnegative real-valued function. Therefore, as shown in (7.0.5), the matrices $[a_{i-j}]$, $[b_{i-j}]$, as well as $[c_{i-j}]$ are positive semidefinite. This is an instance of the Schur product theorem: The Hadamard product of two positive semidefinite matrices is positive semidefinite.

As another example, consider the integral operator

$$K(f) = \int_a^b K(x, y) f(y)\, dy$$

where the kernel $K(x, y)$ is a continuous function on the finite interval $[a, b] \times [a, b]$ and $f \in C[a, b]$. If one has a second kernel $H(x, y)$, then one could consider the (point-wise) product kernel $L(x, y) = K(x, y) H(x, y)$ and the associated integral operator

$$L(f) \equiv \int_a^b L(x, y) f(y)\, dy = \int_a^b K(x, y) H(x, y) f(y)\, dy$$

The linear mapping $f \to K(f)$ is in a natural way a limit of matrix–vector multiplications (approximate the integral as a finite Riemann sum), and many properties of integral operators can be deduced by taking appropriate limits of results known for matrices. The (point-wise) product of integral kernels leads to an integral operator that is, from this point of view, a natural continuous analog of the Hadamard product of matrices.

If the integral kernel $K(x, y)$ has the property that

$$\int_a^b \int_a^b K(x, y) f(x) \bar{f}(y)\, dx\, dy \geq 0$$

for all $f \in C[a, b]$, then $K(x, y)$ is said to be a *positive semidefinite kernel*. It is a classical result (Mercer's theorem) that if $K(x, y)$ is a continuous positive semidefinite kernel on a finite interval $[a, b]$, then there exist positive real numbers $\{\lambda_i\}$ (known as "eigenvalues") and continuous functions $\{\phi_i(x)\}$ (known as "eigenfunctions") such that

$$K(x, y) = \sum_{i=1}^{\infty} \frac{\phi_i(x) \bar{\phi}_i(y)}{\lambda_i} \quad \text{on} \quad [a, b] \times [a, b]$$

and the series converges absolutely and uniformly.

If $K(x, y)$ and $H(x, y)$ are both continuous positive semidefinite kernels on the same finite interval $[a, b]$, then $H(x, y)$ also has an absolutely and uniformly convergent representation

$$H(x, y) = \sum_{i=1}^{\infty} \frac{\psi_i(x) \bar{\psi}_i(y)}{\mu_i} \quad \text{on} \quad [a, b] \times [a, b]$$

with all $\mu_i > 0$. By direct multiplication of the respective series, the (point-wise) product kernel $L(x, y) = K(x, y) H(x, y)$ has the representation

$$L(x, y) = \sum_{i,j=1}^{\infty} \frac{\phi_i(x) \psi_j(x) \bar{\phi}_i(y) \bar{\psi}_j(y)}{\lambda_i \mu_j} \quad \text{on} \quad [a, b] \times [a, b]$$

which also converges absolutely and uniformly. Then

$$\int_a^b \int_a^b L(x, y) f(x) \bar{f}(y)\, dx\, dy = \sum_{i,j=1}^{\infty} \frac{1}{\lambda_i \mu_j} \left| \int_a^b \phi_i(x) \psi_j(x) f(x)\, dx \right|^2 \geq 0$$

so $L(x, y)$ is also positive semidefinite. This is another instance of the Schur product theorem.

Exercise. Show that the Hadamard product of two Hermitian matrices is always Hermitian, but that the usual matrix product of two Hermitian matrices is Hermitian if and only if they commute.

Exercise. Consider the matrices $A = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$. Show that A, B, and $A \circ B$ are positive semidefinite, but that the usual matrix product AB is not positive semidefinite. Show that the eigenvalues of AB are positive, however.

The main reason for introducing the Hadamard product at this point is that it (unlike the usual matrix product) leaves invariant the cone of positive semidefinite matrices and provides yet another analogy between positive semidefinite matrices and nonnegative real numbers. We begin with an observation of independent interest.

Any matrix may be written as a sum, of rank 1 matrices, in which the number of summands is equal to the rank, but for a positive semidefinite matrix the summands may also be chosen to be positive semidefinite.

7.5.2 **Theorem.** If $A \in M_n$ is a positive semidefinite matrix of rank k, then A may be written in the form

$$A = v_1 v_1^* + v_2 v_2^* + \cdots + v_k v_k^*$$

where each $v_i \in \mathbf{C}^n$ and the set $\{v_1, \ldots, v_k\}$ is an orthogonal set of nonzero vectors.

Proof: Use the spectral theorem to write $A = U\Lambda U^*$ and let v_i be $\lambda_i^{1/2}$ times the ith column of U. □

Our main result is often called the *Schur product theorem.*

7.5.3 Theorem. If $A, B \in M_n$ are positive semidefinite matrices, then $A \circ B$ is also positive semidefinite. Moreover, if both A and B are positive definite, then so is $A \circ B$.

Proof: Use (7.5.2) to write $A = v_1 v_1^* + \cdots + v_k v_k^*$ and $B = w_1 w_1^* + \cdots + w_m w_m^*$, where $k = \text{rank } A$ and $m = \text{rank } B$. Observe that

$$A \circ B = \sum_{i,j=1}^{k,m} u_{ij} u_{ij}^*$$

where $u_{ij} = v_i \circ w_j$. Thus, $A \circ B$ is positive semidefinite because it is a sum of (rank 1) positive semidefinite matrices.

If A and B are both positive definite, then $k = m = n$ and the sets $\{v_i\}$ and $\{w_i\}$ are both orthogonal bases of $\mathbf{C}^n$. If $A \circ B$ were singular, there would be some nonzero vector x such that $(A \circ B)x = 0$ and hence

$$x^*(A \circ B)x = \sum_{i,j=1}^{k,m} x^*(u_{ij} u_{ij}^*)x = \sum_{i,j=1}^{k,m} |x^* u_{ij}|^2 = 0$$

But then each term must vanish separately, and hence

$$|x^* u_{ij}|^2 = |x^*(v_i \circ w_j)|^2 = |(x \circ \bar{v}_i)^* w_j|^2 = 0$$

for all i and j. This means that for each i the vector $x \circ \bar{v}_i$ is orthogonal to all the vectors $w_1, w_2, \ldots, w_n$ and therefore $x \circ \bar{v}_i = 0$ for all $i = 1, 2, \ldots, n$. In particular, this implies that $v_i^* x = 0$ for all $i = 1, \ldots$. Since this means that x is orthogonal to all the elements of a basis, we must have $x = 0$. We conclude that $A \circ B$ must be nonsingular. □

Exercise. Let $A, B \in M_n$. Use the proof of (7.5.3) to show that rank $A \circ B \leq$ (rank A)(rank B) whenever A and B are positive semidefinite. In particular, show that if (rank A)(rank B) $< n$, then $A \circ B$ must be singular.

Exercise. Show that the assertions of the previous exercise are still correct when A and B are Hermitian matrices that are not necessarily positive semidefinite.

Exercise. Consider the matrices $A = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ and show that rank $A \circ B$ can be zero even when A and B both have positive rank.

Exercise. Show that if A is positive definite and B is negative definite, then $A \circ B$ is negative definite.

7.5.4 Corollary (Fejer's theorem). Let $A = [a_{ij}] \in M_n$. Then A is positive semidefinite if and only if

$$\sum_{i,j=1}^{n} a_{ij} b_{ij} \geq 0$$

for all positive semidefinite matrices $B = [b_{ij}] \in M_n$.

Proof: Suppose A and B are positive semidefinite, and let $x \in \mathbf{C}^n$ be a vector with all components equal to 1. Then $A \circ B$ is positive semidefinite, and the indicated sum is just $x^*(A \circ B)x$, which much be nonnegative. Conversely, if $\sum a_{ij} b_{ij} \geq 0$ whenever B is positive semidefinite, just set $B = [b_{ij}] \equiv [\bar{x}_i x_j]$ for any given vector $x \in \mathbf{C}^n$. Then B is positive semidefinite and

$$\sum_{i,j=1}^{n} a_{ij} b_{ij} = \sum_{i,j=1}^{n} a_{ij} \bar{x}_i x_j = x^* A x \geq 0$$

Since $x \in \mathbf{C}^n$ was arbitrary, we conclude that A is positive semidefinite. □

7.5.5 Application. Let $D \subset \mathbf{R}^n$ be an open bounded set. The second-order linear differential operator L on $C^2(D)$ given by

$$Lu \equiv \sum_{i,j=1}^{n} a_{ij}(x) \frac{\partial^2 u}{\partial x_i \, \partial x_j} + \sum_{i=1}^{n} b_i(x) \frac{\partial u}{\partial x_i} + c(x)u \qquad (7.5.6)$$

is said to be *elliptic* in D if the matrix $A(x) \equiv [a_{ij}(x)]$ is positive definite for all $x \in D$. Suppose there is some $u \in C^2(D)$ that is continuous on the closure of D and satisfies the equation $Lu \equiv 0$ in D. What can we say about the local maxima or minima of the function u in D? Suppose $y \in D$ is a local minimum for u. Then

$$\left. \frac{\partial u}{\partial x_i} \right|_y = 0 \quad \text{for all} \quad i = 1, 2, \ldots, n$$

and the Hessian matrix

$$\left[\frac{\partial^2 u}{\partial x_i \, \partial x_j} \right]$$

is positive semidefinite at y. Therefore,

$$Lu = 0 = \sum_{i,j=1}^{n} a_{ij} \frac{\partial^2 u}{\partial x_i \, \partial x_j} + cu$$

at the point y, and by Fejer's theorem (7.5.4) the sum involving the second derivatives must be nonnegative. Thus, the term $c(y)u(y)$ must be nonpositive. In particular, if $c(y) < 0$, then it cannot be that $u(y) < 0$. A similar argument shows that $u(y)$ cannot be positive at an interior relative maximum y if $c(y) < 0$. These simple observations are the heart of the following important principle.

7.5.7 **Weak minimum principle.** Let the operator L defined by (7.5.6) be elliptic in D, and suppose that $c(x) < 0$ in D. If $u \in C^2(D)$ satisfies $Lu \equiv 0$ in D, then u cannot have a negative interior relative minimum nor a positive interior relative maximum. If, in addition, u is continuous on the closure of D and u is nonnegative on the boundary of D, then u must be nonnegative everywhere in D.

From the minimum principle follows one of the fundamental uniqueness theorems for partial differential equations:

7.5.8 **Fejer's uniqueness theorem.** Suppose the operator L defined by (7.5.6) is elliptic, assume that $c(x) < 0$ in D, and consider the following boundary value problem:

$Lu \equiv f$ in D, f a given function
$u \equiv g$ on ∂D, g a given function
u is twice continuously differentiable in D
u is continuous on the closure of D

Then there is at most one solution to this problem.

Proof: If u_1 and u_2 were two solutions to this problem, then the function $v \equiv u_1 - u_2$ is a solution to a problem of the same type, but with zero boundary conditions and $Lv \equiv 0$ in D. By the weak minimum principle, v must be nonnegative in D. Applying the same argument to $-v$, we find that v must also be nonpositive in D as well, and hence $v \equiv 0$ in D.

Exercise. Explain how the weak minimum principle and the Fejer uniqueness theorem apply to the partial differential equation $\nabla^2 u - \lambda u = 0$ in $D \subset \mathbf{R}^n$, where λ is a positive real parameter.

A final corollary of the Schur product theorem is easily proved. If $A = [a_{ij}] \in M_n$ is positive semidefinite, then $A \circ A = [a_{ij}^2]$ is also positive semidefinite. By induction, it follows that all the positive integer Hadamard powers $[a_{ij}^k]$ are positive semidefinite for all $k = 1, 2, \ldots$. Since any nonnegative linear combination of positive semidefinite matrices is positive semidefinite (7.1.3), this implies that

$$\begin{aligned} a_0 I + a_1 A + a_2 A \circ A + \cdots + a_m \overbrace{A \circ \cdots \circ A}^{m \text{ times}} &= [a_0 + a_1 a_{ij} + a_2 a_{ij}^2 + \cdots + a_m a_{ij}^m] \\ &= [p(a_{ij})] \end{aligned}$$

is positive semidefinite whenever all $a_i \geq 0$; $p(x) = a_0 + a_1 x + \cdots + a_m x^m$ is a polynomial with nonnegative coefficients. More generally, if

$$f(z) = \sum_{k=0}^{\infty} a_k z^k$$

is an analytic function with all $a_k \geq 0$ and radius of convergence $R > 0$, then a simple limiting argument shows that $[f(a_{ij})] \in M_n$ is positive semidefinite whenever all $|a_{ij}| < R$. Perhaps the simplest example is $f(z) = e^z$, whose power series converges for all $z \in \mathbf{C}$ and whose coefficients are $a_k = 1/k! > 0$. By this argument, $[e^{a_{ij}}]$ is a positive semidefinite matrix whenever $A = [a_{ij}] \in M_n$ is positive semidefinite. This result can be improved; weaker conditions on A are sufficient to ensure that the entrywise exponential of A is positive semidefinite. See [HJ].

7.5.9 **Corollary.** Let $A = [a_{ij}] \in M_n$ be positive semidefinite. Then

(a) The matrix $[a_{ij}^k]$ is positive semidefinite for all $k = 1, 2, \ldots$.
(b) If $f(z) = a_0 + a_1 z + a_2 z^2 + \cdots$ is an analytic function with nonnegative coefficients and radius of convergence $R > 0$, then the matrix $[f(a_{ij})]$ is positive semidefinite if all $|a_{ij}| < R$.

Problems

1. Show that if $H(A)$ (the Hermitian part of A) is positive definite and B is positive definite, then $H(A \circ B)$ is positive definite.

2. If $A = [a_{ij}] \in M_n$ is positive semidefinite, show that the matrix $[|a_{ij}|^2]$ is also positive semidefinite. *Hint:* Consider $A \circ \bar{A}$.

3. If $A = [a_{ij}] \in M_n$ is positive semidefinite, show that the matrix $[e^{e^{a_{ij}} - \lambda}]$ is positive semidefinite for all $\lambda \in \mathbf{R}$.

4. If $A=[a_{ij}]\in M_n$ is positive semidefinite, then the positive integer Hadamard power matrices $A^{(k)}$ and the Hadamard absolute value squared matrix $A\circ\bar{A}$ are always positive semidefinite. But what about the Hadamard absolute value matrix $|A|\equiv[|a_{ij}|]$? (a) Suppose $A\in M_n$ is positive definite. For $n=1,2,3$ use the determinant criterion (7.2.5) to show directly that $|A|$ is positive definite. Obtain the result for positive semidefinite A ($n=1,2,3$ only) by passing to the limit. (b) Use the fact that $f(x)=\cos(x)$ is a positive definite function [or write $\cos(x)=(e^{ix}+e^{-ix})/2$ and compute the quadratic form explicitly] to show that the matrix $A=[\cos(x_i-x_j)]$ is positive semidefinite for all choices of $\{x_1 x_2,\ldots,x_n\}\in\mathbf{R}$ and for all $n=1,2,\ldots$. (c) Let $n=4$, and let $x_1=0$, $x_2=\pi/4$, $x_3=\pi/2$, and $x_4=3\pi/4$. Explicitly compute the (necessarily positive semidefinite) matrix A in (b) in this case, and observe that it is a Toeplitz matrix. Compute $|A|$ and $\det|A|$, and show that $|A|$ cannot be positive semidefinite.

5. Consider the matrix $|A|$ in Problem 4, and show that $B\equiv|A|\circ|A|$ is an example of a positive semidefinite matrix whose nonnegative "Hadamard square root" is not positive semidefinite. Contrast this with the situation for the ordinary square root $B^{1/2}$.

6. Consider the matrix $A\in M_4$ given by

$$A=\begin{bmatrix} 10 & 3 & -2 & 1\\ 3 & 10 & 0 & 9\\ -2 & 0 & 10 & 4\\ 1 & 9 & 4 & 10\end{bmatrix}$$

Show that A is positive definite but that $|A|$ is not positive semidefinite.

7. Let $K(x,y)$ be a continuous integral kernel on the finite interval $[a,b]$. Show that $K(x,y)$ is a positive semidefinite kernel if and only if the matrix $[K(x_i,x_j)]\in M_n$ is positive semidefinite for all choices of the points $\{x_i\}_{i=1}^n\subset[a,b]$ and all $n=1,2,\ldots$. *Hint:* To show that the matrix condition is sufficient, consider a Riemann sum approximation to the integral

$$\int_a^b\int_a^b K(x,y)f(x)\bar{f}(y)\,dx\,dy\cong\sum_{i,j=1}^n K(x_i,x_j)f(x_i)\bar{f}(x_j)\,\Delta x_i\,\Delta x_j$$

To show that the matrix condition is necessary, consider a function

$$f(x)\equiv\sum_{i=1}^n a_i\delta_\epsilon(x-x_i)$$

where $\delta_\epsilon(x)$ is an "approximate delta function," which is continuous and nonnegative, vanishes identically outside the interval $[-\epsilon,\epsilon]$, and satisfies

$$\int_{-\infty}^{\infty}\delta_\epsilon(x)\,dx=1$$

Now let $\epsilon\to 0$.

8. Use Problem 7 and the Schur product theorem to show that the (point-wise) product of positive semidefinite integral kernels is positive semidefinite. This line of argument is relatively elementary and does not require Mercer's theorem from the theory of integral equations.

9. Show that a function $\phi\in C(\mathbf{R})$ is a positive definite function [Problem 8 in Section (7.1)] if and only if $K(x,y)\equiv\phi(x-y)$ is a positive semidefinite integral kernel.

10. Show that the product $(\phi_1\phi_2)(x)$ of two positive definite functions $\phi_1(x),\phi_2(x)$ is a positive definite function.

11. Explain why all the functions

(a) $\dfrac{\sin(Tx)}{Tx}=\dfrac{1}{2T}\displaystyle\int_{-T}^{T}e^{itx}\,dt,\qquad T>0$

(b) $e^{-x^2}=\dfrac{1}{2\sqrt{\pi}}\displaystyle\int_{-\infty}^{\infty}e^{-t^2/2}e^{itx}\,dt$

(c) $e^{-|x|}=\dfrac{1}{\pi}\displaystyle\int_{-\infty}^{\infty}\dfrac{e^{itx}}{1+t^2}\,dt$

as well as all their mutual products are positive definite functions.

12. Use Problem 11(c) to give an alternate proof that the matrix in Problem 12 of Section (7.2) is positive definite.

13. If $A=[a_{ij}]\in M_n$ is positive semidefinite, show that the matrix $[a_{ij}/(i+j)]$ is also positive semidefinite. *Hint:* Problem 17 of Section (7.1).

14. Let $A\in M_n$ be positive semidefinite. Show that $x\in\mathbf{C}^n$ satisfies $x^*Ax=0$ if and only if $Ax=0$. If A is merely Hermitian, show by example that it is possible to have $x^*Ax=0$ and $Ax\neq 0$. *Hint:* Write $A=U\Lambda U^*$, so $x^*Ax=0$ if and only if $\sum\lambda_i|z_i|^2=0$, where $z=U^*x$.

15. A convex cone (with vertex at the origin 0) is a convex set S such that the ray $\{\lambda x:\lambda\geq 0\}\subset S$ for all $x\in S$. A ray $\{\lambda x:\lambda\geq 0\}$ in a convex cone S is an *extreme ray* if $x=\alpha y+(1-\alpha)z$ for $0<\alpha<1$ and $y,z\in S$ only if

both y and z lie on the ray; equivalently, a ray of a convex cone is extreme if it can be deleted from the cone and the resulting cone is still convex. Show that a ray $\{\lambda A : \lambda \geq 0\}$ in the convex cone of positive semidefinite matrices in M_n is an extreme ray if and only if A has rank 1. Theorem (7.5.2) then says that every positive semidefinite matrix is a convex combination of matrices that lie on extreme rays. *Hint:* (a) If $x \in \mathbf{C}^n$ is nonzero, and if $xx^* = \alpha A + (1-\alpha)B$ for some $\alpha \in (0,1)$ and positive semidefinite matrices $A, B \in M_n$, let $\{x_1, \ldots, x_n\} \subset \mathbf{C}^n$ be an orthonormal set such that $x^*x_k = 0$ for $k = 2, \ldots, n$. Then $0 = x_k^* A x_k = x_k^* B x_k$, and hence each x_k is in the null space of both A and B by Problem 14. Conclude that both A and B have rank 1 and are positive scalar multiples of xx^*. (b) If $A \in M_n$ is positive semidefinite and has rank $k \geq 2$, use (7.5.2) to write $A = B + C$, where $B = vv^*$, $v \neq 0$, rank $C \geq 1$, and $Cv = 0$. Conclude that C is not a scalar multiple of B and hence A does not lie on an extreme ray.

7.6 Congruence: products and simultaneous diagonalization

Unlike multiplication of positive real numbers, the usual matrix multiplication does not always preserve positive definiteness. The product of two Hermitian matrices may not even be Hermitian (it is Hermitian only when they commute), and the quadratic form generated by the product may not be nonnegative. The particular focus of this section is on positive definite matrices; for more general results about Hermitian matrices see Section (4.5).

7.6.1 **Example.** Let $A = \begin{bmatrix} 5 & -2 \\ -2 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$ so that A and B are positive definite. But $AB = \begin{bmatrix} 8 & 3 \\ -3 & -1 \end{bmatrix}$ is not symmetric, and not even $H(AB) = \begin{bmatrix} 8 & 0 \\ 0 & -1 \end{bmatrix}$ is positive definite.

At least one positivity property is retained by the usual matrix product of positive definite matrices, however. Our discussion will illustrate some useful techniques for dealing with products and sums of matrices.

7.6.2 **Definition.** Recall that two matrices $A, B \in M_n$ are **congruent* if there exists a nonsingular matrix $C \in M_n$ such that $B = C^*AC$.

Note that, like similarity, *congruence is an equivalence relation. Sometimes the term *conjunctive* is used in the complex case to distinguish it from real congruence.

7.6.3 **Theorem.** The product of a positive definite matrix $A \in M_n$ and a Hermitian matrix $B \in M_n$ is a diagonalizable matrix, all of whose eigenvalues are real. The matrix AB has the same number of positive, negative, and zero eigenvalues as B. Furthermore, any diagonalizable matrix with real eigenvalues is the product of a positive definite matrix and a Hermitian matrix.

Proof: For the first part, note that $A^{-1/2}ABA^{1/2} = A^{1/2}BA^{1/2}$, so the latter matrix is similar to AB and hence has exactly the same eigenvalues. Since $A^{1/2}$ is Hermitian, the matrix $A^{1/2}BA^{1/2}$ is congruent to B. Thus, by Sylvester's law of inertia (4.5.8), the eigenvalues of B have the same set of signs as those of $A^{1/2}BA^{1/2}$, and hence of AB. Moreover, since $A^{1/2}BA^{1/2}$ is Hermitian, it is diagonalizable, and hence AB must also be diagonalizable. For the last assertion, suppose $C \in M_n$ is diagonalizable and has only real eigenvalues: $C = SDS^{-1}$, with D a real diagonal matrix. Then $C = S(S^*S^{*-1})DS^{-1} = (SS^*)(S^{-1*}DS^{-1}) = AB$ where $A \equiv SS^*$ is positive definite and $B \equiv S^{-1*}DS^{-1}$ is Hermitian. □

Simultaneous diagonalizability of two matrices by similarity is a rare event, requiring the strong joint assumption of commutativity. Simultaneous diagonalization of two Hermitian matrices by joint *congruence, however, requires much less. Simultaneous diagonalization by *congruence corresponds to transforming two Hermitian quadratic forms into a linear combination of squares by a single linear change of variables. The following result is classical; for a generalization see (4.5.15).

7.6.4 **Theorem.** Let $A, B \in M_n$ be two Hermitian matrices and suppose that there is a real linear combination of A and B that is positive definite. Then there exists a nonsingular matrix $C \in M_n$ such that both C^*AC and C^*BC are diagonal.

Proof: Suppose that $P = \alpha A + \beta B$ is positive definite for some $\alpha, \beta \in \mathbf{R}$. At least one of α and β must be nonzero, so we may assume that $\beta \neq 0$. But since $B = \beta^{-1}(P - \alpha A)$, if we can show that A and P are simultaneously diagonalizable by *congruence, then it will follow that A and B are also. By (7.2.7) we know that P is *congruent to the identity, so there is some nonsingular $C_1 \in M_n$ such that $C_1^* P C_1 = I$. Since $C_1^* A C_1$ is Hermitian, there exists a unitary matrix U such that $U^* C_1^* A C_1 U = D$ is diagonal. Letting $C \equiv C_1 U$, we have $C^*PC = I$ and $C^*AC = D$ so that $C^*BC = \beta^{-1}(I - \alpha D)$ is diagonal. □

The most common application of this result is to the classical situation in mechanics, in which two real symmetric quadratic forms are given, one of which is positive definite.

7.6.5 Corollary. If $A \in M_n$ is positive definite and $B \in M_n$ is Hermitian, then there exists a nonsingular matrix $C \in M_n$ such that C^*BC is diagonal and $C^*AC = I$.

Exercise. Find a single change of variables such that both quadratic forms $5x^2 - 2xy + y^2$ and $x^2 + 2xy - y^2$ are weighted sums of squares.

There is an analogous result for a pair of matrices, one of which is positive definite and the other (complex) symmetric. This result is also generalized in (4.5.15).

7.6.6 Theorem. If $A \in M_n$ is positive definite and $B \in M_n$ is a symmetric complex matrix, then there is a nonsingular matrix C such that C^*AC and C^TBC are both diagonal.

Proof: Choose a nonsingular matrix $C_1 \in M_n$ such that $C_1^*AC_1 = I$. Then $C_1^TBC_1$ is symmetric, so by Takagi's factorization (4.4.4) there is a unitary matrix U such that $U^T(C_1^TBC_1)U = D$, where D is diagonal. Then $U^*C_1^*AC_1U = I$, too, so we may take $C \equiv C_1U$. □

This result has applications to complex function theory; the Grunsky inequalities for univalent functions are inequalities between quadratic forms generated by a positive definite Hermitian matrix and a complex symmetric matrix.

The following result is an immediate application of (7.6.5).

7.6.7 Theorem. The function $f(A) = \log \det A$ is a strictly concave function on the convex set of positive definite Hermitian matrices in M_n.

Proof: For any two given positive definite matrices $A, B \in M_n$ we must show that

$$f(\alpha A + (1-\alpha)B) \geq \alpha f(A) + (1-\alpha) f(B) \tag{7.6.8}$$

for all $\alpha \in (0,1)$, with equality if and only if $A = B$. Use (7.6.5) to write $A = CIC^*$ and $B = C\Lambda C^*$ for some nonsingular $C \in M_n$ and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ with all $\lambda_i > 0$. Then

$$\begin{aligned} f(\alpha A + (1-\alpha)B) &= f(C[\alpha I + (1-\alpha)\Lambda]C^*) = f(CC^*) + f(\alpha I + (1-\alpha)\Lambda) \\ &= f(A) + f(\alpha I + (1-\alpha)\Lambda) \end{aligned}$$

and

$$\begin{aligned} \alpha f(A) + (1-\alpha) f(B) &= \alpha f(A) + (1-\alpha) f(C\Lambda C^*) \\ &= \alpha f(A) + (1-\alpha)[f(CC^*) + f(\Lambda)] \\ &= \alpha f(A) + (1-\alpha) f(A) + (1-\alpha) f(\Lambda) \\ &= f(A) + (1-\alpha) f(\Lambda) \end{aligned}$$

Thus, it suffices to show that $f(\alpha I + (1-\alpha)\Lambda) \geq (1-\alpha) f(\Lambda)$ for all $\alpha \in (0,1)$ for any diagonal matrix Λ with positive diagonal entries. But this follows easily from the strict concavity of the logarithm function itself since

$$\begin{aligned} f(\alpha I + (1-\alpha)\Lambda) &= \log \prod_{i=1}^n [\alpha + (1-\alpha)\lambda_i] = \sum_{i=1}^n \log[\alpha + (1-\alpha)\lambda_i] \\ &\geq \sum_{i=1}^n [\alpha \log 1 + (1-\alpha) \log \lambda_i] \\ &= (1-\alpha) \sum_{i=1}^n \log \lambda_i = (1-\alpha) \log \prod_{i=1}^n \lambda_i \\ &= (1-\alpha) \log \det \Lambda = (1-\alpha) f(\Lambda) \end{aligned}$$

Equality holds in this inequality if and only if every $\lambda_i = 1$, which can happen if and only if $\Lambda = I$ and $B = CIC^* = A$. □

Theorem (7.6.7) is often used in the following form, which is obtained by exponentiating the inequality (7.6.8). It gives a quantitative expression for the fact that a convex combination of positive definite matrices is positive definite, and hence must be nonsingular.

7.6.9 Corollary. Let $A, B \in M_n$ be positive definite and let $0 < \alpha < 1$. Then

$$\det[\alpha A + (1-\alpha)B] \geq [\det A]^{\alpha}[\det B]^{1-\alpha}$$

with equality if and only if $A = B$.

Problems

1. Suppose $A \in M_n$ satisfies $A^* = S^{-1}AS$ with $S \in M_n$ positive definite. Show that A is diagonalizable and all the eigenvalues of A are real. *Hint:* Consider $AS = SA^*$. Show that AS is Hermitian and use (7.6.3).

2. Show that $f(A) = \operatorname{tr} A^{-1}$ is a strictly convex function on the positive definite matrices. *Hint:* The proof of (7.6.7).

3. How does (7.6.3) generalize if $A \in M_n$ is positive *semi*definite? Show that the eigenvalues of AB are still real and that AB has no more positive eigenvalues and no more negative eigenvalues than B has, but that it may have more zero eigenvalues.

4. To what extent does (7.6.3) generalize if $B \in M_n$ is not Hermitian?

5. Show by example that Hermitian matrices might be simultaneously diagonalized by congruence without the hypothesis of (7.6.4) being satisfied.

6. Let $A, B \in M_2$ be given Hermitian matrices. What are all the possibilities for the signs of real parts of the eigenvalues of AB in terms of those of A and B? Can you generalize this to M_n?

7. Let $A, B \in M_n$ be Hermitian with A positive definite. Use (7.6.5) to show that $A+B$ is positive definite if and only if every eigenvalue of $A^{-1}B$ is greater than -1. *Hint:* $A+B = A(I+A^{-1}B)$.

8. Let $H \in M_n$ be Hermitian, and write $H = A + iB$ with $A, B \in M_n(\mathbf{R})$. Verify that A is symmetric and B is skew-symmetric, so the eigenvalues of B are pure imaginary and occur in conjugate pairs. Show that H is positive definite if and only if A is positive definite and every eigenvalue of $iA^{-1}B$ is greater than -1. *Hint:* Use the fact that $x^*Hx = x^*Ax$ for all $x \in \mathbf{R}^n$. Use Problem 7. If A is positive definite, show that if λ is an eigenvalue of $iA^{-1}B$, then so is $-\lambda$. Conclude that H is positive definite if and only if A is positive definite and every eigenvalue of $iA^{-1}B$ lies in the interval $(-1, 1)$, and the eigenvalues of $iA^{-1}B$ occur in pairs $\{-\lambda, \lambda\}$. Conclude that $0 \leq \det iA^{-1}B < 1$ and hence that $\det B < \det A$, an inequality of H. P. Robertson. Now write $H = A + iB = A(I + iA^{-1}B)$ and show that if H is positive definite, then $\det H = \det A \det(I + iA^{-1}B)$ and $0 < \det(I + iA^{-1}B) < 1$. Conclude that if H is positive definite, then $\det H \leq \det A$, an inequality of O. Taussky. See (7.8.7) and Problem 7 of Section (7.8) for a version of these inequalities that is valid for general complex matrices $H \in M_n$.

9. In Theorem (4.1.7) we found that a matrix $A \in M_n$ is the product of two Hermitian matrices if and only if A is similar to a real matrix. Use (7.6.3) to show that $A \in M_n$ is the product of two positive definite Hermitian matrices if and only if A is diagonalizable and has positive eigenvalues. *Hint:* For the converse, consider $A = S\Lambda S^{-1} = SS^*(S^{-1})^*\Lambda S^{-1}$.

10. If $A, B \in M_n$ are positive definite, then we know that the product AB is positive definite if and only if AB is Hermitian. Show that the same is true for products of three positive definite matrices; that is, if $A, B, C \in M_n$ are positive definite, then the product $S = ABC$ is positive definite if and only if it is Hermitian. *Hint:* Write $S = (AB)C = EC$, where E has n positive eigenvalues by Problem 9. Use (7.6.3) to show that $E = SC^{-1}$ has the same number of positive eigenvalues as S if S is Hermitian.

11. Provide the details for the following alternate proof of the result in Problem 10: Let $S(\alpha) \equiv [(1-\alpha)C + \alpha A]BC$ for $0 \leq \alpha \leq 1$. If $S = S(1)$ is Hermitian, then all $S(\alpha)$ are Hermitian because $S(0) = CBC$ is automatically Hermitian. Argue that all $S(\alpha)$ are nonsingular because $(1-\alpha)C + \alpha A$ is nonsingular. The eigenvalues of $S(\alpha)$ depend continuously on α, all the eigenvalues are positive for $\alpha = 0$, and no eigenvalue vanishes since all $S(\alpha)$ are nonsingular. Conclude that all eigenvalues of $S(1)$ are positive.

Further Reading. For further results about products of matrices from various positivity classes and references to earlier results about products of several positive definite matrices, see C. S. Ballantine and C. R. Johnson, "Accretive Matrix Products," *Lin. Multilin. Alg.* 3 (1975), 169–185.

7.7 The positive semidefinite ordering

Because Hermitian matrices are generalizations of real numbers and positive definite matrices are generalizations of positive real numbers, it is natural to ask whether there is a good notion of inequality or (partial) order among Hermitian matrices.

7.7.1 **Definition.** Let $A, B \in M_n$ be Hermitian matrices. We write $A \succcurlyeq B$ if the matrix $A - B$ is positive semidefinite. Similarly, $A \succ B$ means that $A - B$ is positive definite.

Exercise. Show that the above notion of inequality is consistent with the notion of equality of matrices. That is, show that $A \succcurlyeq B$ and $B \succcurlyeq A$ imply that $A = B$.

Exercise. Show that the relation $\succcurlyeq$ is transitive and reflexive, but that it is not a total order; that is, there exist Hermitian matrices $A, B \in M_n$ such that neither $A \succcurlyeq B$ nor $B \succcurlyeq A$. Such a relation is called a *partial order.*

A partial order on a real linear space is often defined by identifying some special closed convex cone and saying that one element is greater than or equal to the other if their difference lies in the special cone. In this case, the set of Hermitian n-by-n matrices is the real linear space and the set of positive semidefinite matrices is the closed convex cone. This is clearly a generalization of the familiar case in which $\mathbf{R}$ itself is the real linear space and the nonnegative real numbers are the closed convex cone: This gives the "usual" (total) order (not just a partial order) on $\mathbf{R}$.

Various other notions of "inequality" among matrices, most notably component-wise domination for real matrices, can be defined in a similar manner: Identify a cone of matrices generalizing the nonnegative real numbers and say that A is "greater than or equal to" B if their difference $A-B$ lies in the cone. Generally, different such notions of "inequality" can be distinguished by the context, but their utility hinges upon how far the analogy with real numbers extends and how strongly the notion relates to other inequalities (such as those among eigenvalues, determinants, etc.).

Notice that A is positive semidefinite if and only if $A \succcurlyeq 0$ and A is positive definite if and only if $A \succ 0$, where 0 is the zero matrix of the same size as A.

Exercise. Show by example that the positive semidefinite partial ordering differs from the total ordering of real numbers in the following way: If $A \succcurlyeq B$ and if A is not equal to B, it does not follow that $A \succ B$.

We next illustrate some of the properties of the positive semidefinite ordering, each of which may be thought of as generalizing the usual ordering of the reals. The analogy is generally quite strong.

7.7.2 **Observation.** If $A, B \in M_n$ are Hermitian, then

$$A \succcurlyeq B \quad \text{implies} \quad T^*AT \succcurlyeq T^*BT$$

for all $T \in M_{n,m}$; if $m \le n$, we also have

$$A \succ B \quad \text{implies} \quad T^*AT \succ T^*BT$$

whenever $T \in M_{n,m}$ has rank m.

Proof: If $A-B$ is positive semidefinite, then $y^*(A-B)y \ge 0$ for all $y \in \mathbf{C}^n$. Thus, $x^*(T^*AT-T^*BT)x = (Tx)^*(A-B)(Tx) \ge 0$ for all $x \in \mathbf{C}^n$ which, in turn, means that T^*AT-T^*BT is positive semidefinite and therefore that $T^*AT \succcurlyeq T^*BT$. Note that this generalizes (7.1.6) and the proof is essentially the same.

Exercise. Verify the second statement above to complete the proof. □

7.7.3 **Theorem.** Let $A, B \in M_n$ be Hermitian matrices, and suppose A is positive definite and B is positive semidefinite. Then $A \succcurlyeq B$ if and only if $\rho(BA^{-1}) \le 1$, and $A \succ B$ if and only if $\rho(BA^{-1}) < 1$.

Proof: By (7.6.5) we can find a nonsingular $C \in M_n$ such that $A = CIC^*$ and $B = CDC^*$, where $D = \operatorname{diag}(d_1, d_2, \ldots, d_n)$ is diagonal. Then $A \succcurlyeq B$ if and only if $C[I-D]C^* \ge 0$, which is the case if and only if $d_i \le 1$ for all $i = 1, 2, \ldots$. But since $BA^{-1} = CDC^*C^{*-1}C^{-1} = CDC^{-1}$, the eigenvalues of BA^{-1} are precisely $d_1, d_2, \ldots, d_n$ [which are all nonnegative by (7.6.3)], and all $d_i \le 1$ if and only if $\rho(BA^{-1}) \le 1$. The last assertion follows from a careful examination of the inequalities just used. □

7.7.4 **Corollary.** If $A, B \in M_n$ are positive definite, then

(a) $A \succcurlyeq B$ if and only if $B^{-1} \succcurlyeq A^{-1}$;
(b) If $A \succcurlyeq B$, then $\det A \ge \det B$ and $\operatorname{tr} A \ge \operatorname{tr} B$; and
(c) More generally, if $A \succcurlyeq B$, then $\lambda_k(A) \ge \lambda_k(B)$ for all $k = 1, 2, \ldots, n$ if the respective eigenvalues of A and B are arranged in the same (increasing or decreasing) order.

Proof: We know that $A \succcurlyeq B$ if and only if $\rho(BA^{-1}) \le 1$. But $\rho(BA^{-1}) = \rho(A^{-1}B)$, and (7.7.3) says that $\rho(A^{-1}B) \le 1$ if and only if $B^{-1} \succcurlyeq A^{-1}$. If $A \succcurlyeq B$, then $\rho(BA^{-1}) \le 1$, and since all the eigenvalues of BA^{-1} are nonnegative by (7.6.3), we know they must lie in the interval $(0, 1]$. But then their product is at most 1, so $\det(BA^{-1}) \le 1$ and hence $\det A \ge \det B$. In the proof of (7.7.3) we found that $A = CC^*$ and $B = CDC^*$ with $C = [c_{ij}] \in M_n$, $D = \operatorname{diag}(d_1, d_2, \ldots, d_n) \in M_n$, and $0 \le d_i \le 1$ for all $i = 1, 2, \ldots, n$. One computes easily that

$$\operatorname{tr} A = \operatorname{tr} CC^* = \sum_{i,j=1}^{n} |c_{ij}|^2$$

and

$$\operatorname{tr} B = \operatorname{tr} CDC^* = \operatorname{tr} DC^*C = \sum_{i,j=1}^{n} d_i |c_{ij}|^2$$

$$\le \sum_{i,j=1}^{n} |c_{ij}|^2 = \operatorname{tr} A$$

The last assertion (which implies the determinant and trace inequalities, for which we have given independent proofs) follows immediately from

the Courant–Fischer variational characterization of the ordered eigenvalues of a Hermitian matrix and is covered by Corollary (4.3.3). □

Exercise. If $A > B > 0$, show that $\det A > \det B$ and $\operatorname{tr} A > \operatorname{tr} B$.

The form for the inverse of a partitioned matrix (0.7.3), when specialized to the case of Hermitian matrices, yields the following useful formula:

$$\begin{bmatrix} A & B \\ B^* & C \end{bmatrix}^{-1} = \begin{bmatrix} (A-BC^{-1}B^*)^{-1} & A^{-1}B(B^*A^{-1}B-C)^{-1} \\ (B^*A^{-1}B-C)^{-1}B^*A^{-1} & (C-B^*A^{-1}B)^{-1} \end{bmatrix} \quad (7.7.5)$$

In this formula we assume that A and C are square and that the appropriate matrices are nonsingular.

If the matrix $\left[\begin{smallmatrix} A & B \\ B^* & C \end{smallmatrix}\right]$ is positive definite, then $\left[\begin{smallmatrix} A & B \\ B^* & C \end{smallmatrix}\right]^{-1}$ exists and is positive definite. It then follows from (7.7.5) and (7.1.2) that $(A-BC^{-1}B^*)^{-1}$ and $A-BC^{-1}B^*$ are positive definite. Similarly, $C-B^*A^{-1}B$, A, and C are positive definite. Thus, if the partitioned Hermitian matrix $\left[\begin{smallmatrix} A & B \\ B^* & C \end{smallmatrix}\right]$ is positive definite, we have

$$A > 0, \qquad C > 0, \qquad A > BC^{-1}B^*, \quad \text{and} \quad C > B^*A^{-1}B$$

7.7.6 Theorem. Suppose that a Hermitian matrix is partitioned as

$$\begin{bmatrix} A & B \\ B^* & C \end{bmatrix}$$

where A and C are square. This matrix is positive definite if and only if A is positive definite and $C > B^*A^{-1}B$. Furthermore, this condition is equivalent to having $\rho(B^*A^{-1}BC^{-1}) < 1$.

Proof: The necessity of the two conditions has been noted above. For sufficiency, suppose that A is positive definite and $C > B^*A^{-1}B$, and calculate, for $X = -A^{-1}B$,

$$\begin{bmatrix} I & 0 \\ X^* & I \end{bmatrix}\begin{bmatrix} A & B \\ B^* & C \end{bmatrix}\begin{bmatrix} I & X \\ 0 & I \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & C-B^*A^{-1}B \end{bmatrix}$$

Since the right-hand side is positive definite, the positive definiteness of

$$\begin{bmatrix} A & B \\ B^* & C \end{bmatrix}$$

follows from the exhibited congruence and (7.1.6) or (7.7.2). The last assertion follows from (7.7.3) applied to the inequality $C > B^*A^{-1}B$. □

Exercise. If $\left[\begin{smallmatrix} A & B \\ B^* & C \end{smallmatrix}\right] > 0$, show that $\det C > \det B^*A^{-1}B$ and $\det A > \det BC^{-1}B^*$. What does this say when $B \in M_{n,1}$? Show that $\det A \det C \geq |\det B|^2$ if B is square.

Exercise. Suppose that $A \in M_n$, $C \in M_m$, and $B \in M_{n,m}$, and suppose that both A and C are positive definite. Show that $\left[\begin{smallmatrix} A & B \\ B^* & C \end{smallmatrix}\right] \succcurlyeq 0$ if and only if $\rho(B^*A^{-1}BC^{-1}) \leq 1$.

The partitioned positive definite matrix in (7.7.6) is related to some bilinear inequalities arising in complex function theory and harmonic analysis that share some of the properties of the positive definite partial order.

7.7.7 Theorem. Let $A \in M_n$ and $C \in M_m$ be positive definite, and let $B \in M_{n,m}$. The following are equivalent:

(a) $(x^*Ax)(y^*Cy) \geq |x^*By|^2$ for all $x \in \mathbf{C}^n$ and all $y \in \mathbf{C}^m$
(b) $x^*Ax + y^*Cy \geq 2|x^*By|$ for all $x \in \mathbf{C}^n$ and all $y \in \mathbf{C}^m$
(c) $\rho(B^*A^{-1}BC^{-1}) \leq 1$
(d) $\left[\begin{smallmatrix} A & B \\ B^* & C \end{smallmatrix}\right] \succcurlyeq 0$

Proof: We shall show that (a) implies (b) implies (c) implies (a); we already know that (c) and (d) are equivalent. If (a) holds, then by the arithmetic-geometric mean inequality we have

$$\tfrac{1}{2}(x^*Ax + y^*Cy) \geq (x^*Ax)^{1/2}(y^*Cy)^{1/2} \geq |x^*By|$$

so (b) follows. If we assume (b), then

$$x^*Ax + y^*Cy = (A^{1/2}x)^*(A^{1/2}x) + (C^{1/2}y)^*(C^{1/2}y) \geq 2|x^*By|$$

and hence for every $x \in \mathbf{C}^n$ and every $y \in \mathbf{C}^m$ we have

$$x^*x + y^*y \geq 2|(A^{-1/2}x)^*B(C^{1/2}y)| = 2|x^*A^{-1/2}BC^{-1/2}y|$$

If we set $x \equiv A^{-1/2}BC^{-1/2}y$ in this inequality, we obtain

$$y^*C^{-1/2}B^*A^{-1}BC^{-1/2}y + y^*y \geq 2|y^*C^{-1/2}B^*A^{-1}BC^{-1/2}y|$$

Since the matrix $C^{-1/2}B^*A^{-1/2}BC^{-1/2}$ is positive semidefinite, this is equivalent to

$$y^*y \geq y^*C^{-1/2}B^*A^{-1}BC^{-1/2}y \quad \text{for all} \quad y \in \mathbf{C}^m$$

If y is chosen to be an eigenvector of $C^{-1/2}B^*A^{-1}BC^{-1/2}$, this inequality says that the (necessarily nonnegative) associated eigenvalue is bounded by 1, so we conclude that the spectral radius is at most 1; that is,

$1 \geq \rho(C^{-1/2}B^*A^{-1}BC^{-1/2}) = \rho(B^*A^{-1}BC^{-1})$ and (c) follows. Finally, if (c) holds, then for any $x \in \mathbf{C}^n$ and any $y \in \mathbf{C}^m$ we have

$$|x^*(A^{-1/2}BC^{-1/2}y)|^2 \leq \|x\|_2^2 \|A^{-1/2}BC^{-1/2}y\|_2^2$$
$$= (x^*x)(y^*C^{-1/2}B^*A^{-1}BC^{-1/2}y) \leq (x^*x)(y^*y)$$

where $\|x\|_2 \equiv (x^*x)^{1/2}$ is the Euclidean norm. If we now make the substitution $x \to A^{1/2}x$ and $y \to C^{1/2}y$, we obtain

$$|x^*By|^2 \leq (x^*Ax)(y^*Cy) \quad \text{for all} \quad x \in \mathbf{C}^n \text{ and all } y \in \mathbf{C}^m \qquad \square$$

For another sort of inequality that stems from (7.7.5), we consider two possible operations that might be applied to a positive definite matrix: inversion and extraction of a principal submatrix based upon a prescribed set of indices. We know that both operations preserve positive definiteness, but is there any relation between the result of applying these operations in the two possible orders? The intriguing fact is that these two operations "commute except for an inequality."

7.7.8 **Theorem.** Suppose that $P \in M_n$ is positive definite, and let $S \subset \{1, 2, \ldots, n\}$ be an index set. Then

$$P^{-1}(S) \geq [P(S)]^{-1}$$

where the left-hand side of this inequality is the principal submatrix of P^{-1} determined by deletion of the rows and columns indicated by S, while the right-hand side is the inverse of the corresponding submatrix of P.

Proof: Since the set of positive definite matrices is closed under permutation congruence, we may assume that

$$P = \begin{bmatrix} A & B \\ B^* & C \end{bmatrix}$$

and that $P(S) = A$. Then, $P^{-1}(S) = (A - BC^{-1}B^*)^{-1}$ and $[P(S)]^{-1} = A^{-1}$. Since $C > 0$ (because $P > 0$), we have $BC^{-1}B^* \geq 0$ and

$$A \geq A - BC^{-1}B^* \geq 0$$

The asserted inequality then follows from (7.7.4a). $\square$

Theorem (7.7.8) may be paraphrased "The inverse of a principal submatrix is less than or equal to the corresponding principal submatrix of the inverse" for a positive definite matrix.

One application of (7.7.8) is to a particular choice of a principal submatrix that produces the Hadamard product from the Kronecker product (see [HJ]). If $A, B \in M_n$, and if $S = \{1, n+2, 2n+3, 3n+4, \ldots, n^2\}$, then $A \circ B = (A \otimes B)(S)$. If A and B are invertible, then $A \otimes B$ is invertible and $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$. Thus, if A and B are positive definite, and if (7.7.8) is applied to $P = A \otimes B$, we obtain

$$A^{-1} \circ B^{-1} = (A^{-1} \otimes B^{-1})(S) = (A \otimes B)^{-1}(S) \geq [(A \otimes B)(S)]^{-1} = (A \circ B)^{-1}$$

If we take $B = A$, this says that $A^{-1} \circ A^{-1} \geq (A \circ A)^{-1}$. But if we take $B = A^{-1}$, this says that $A^{-1} \circ A \geq (A \circ A^{-1})^{-1} = (A^{-1} \circ A)^{-1}$ whenever A is positive definite.

The last inequality says that $A^{-1} \circ A$ dominates its own inverse. What does this say about $A^{-1} \circ A$? If C is a positive definite matrix with $C = U\Lambda U^*$, $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$, and all $\lambda_i > 0$, then $C \geq C^{-1}$ if and only if all $\lambda_i \geq 1$, and hence $C \geq I \geq C^{-1}$. We may summarize these observations as

7.7.9 **Theorem.** Let $A, B \in M_n$ be positive definite. Then

(a) $A^{-1} \circ B^{-1} \geq (A \circ B)^{-1}$;
(b) $A^{-1} \circ A^{-1} \geq (A \circ A)^{-1}$; and
(c) $A^{-1} \circ A \geq I \geq (A^{-1} \circ A)^{-1}$.

Since $A^{-1}A = I$, the first part of (c) says that $A^{-1} \circ A \geq A^{-1}A$; that is, Hadamard multiplication dominates ordinary multiplication in this case.

Problems

1. In general, if $A, B \in M_n$ are Hermitian and $A \geq B$, show that if $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are the ordered eigenvalues of A and if $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_n$ are the ordered eigenvalues of B, then $\lambda_i \geq \mu_i$, $i = 1, 2, \ldots, n$. Show by example, however, that the converse is not always valid.

2. If $A_1, A_2, B_1, B_2 \in M_n$ are all Hermitian, show that if $A_1 \geq B_1$ and $A_2 \geq B_2$, then $A_1 + A_2 \geq B_1 + B_2$.

3. Let $A, B, C \in M_n$ be Hermitian; assume that $A \geq B$ and $C \geq 0$. Show that $A \circ C \geq B \circ C$.

4. Let $A, B, C, D \in M_n$ be Hermitian and suppose that $A \geq B \geq 0$ and $C \geq D \geq 0$. Use the previous problem to show that $A \circ C \geq B \circ D \geq 0$.

5. If $A, B \in M_n$ are Hermitian matrices such that $A \geq B$, and if $J \subset \{1, 2, \ldots, n\}$ is any index set, then show that $A(J) \geq B(J)$.

6. Show that (7.7.6) generalizes (7.2.5) for $n = 2$.

7. What does (7.7.6) say if $C \in M_1$? How does one augment a positive definite matrix with one row and column and preserve positive definiteness?

8. Show that the inequality of (7.7.8) is strict if and only if $P(S, S')$ has full row rank and equality holds precisely when $P(S, S')=0$. Here, $P(S, S')$ is the submatrix of P resulting from deletion of the rows indicated by S and the columns indicated by S'. *Hint:* Show that $\operatorname{rank}[P^{-1}(S)-P(S)^{-1}]=\operatorname{rank} P(S, S')$.

9. If $A\in M_n$ is Hermitian, show that $I\succcurlyeq A$ if and only if all the eigenvalues of A are less than or equal to 1.

10. Use (7.7.7) to give an alternate solution to Problem 11 of Section (7.4). *Hint:* Show that $\begin{bmatrix} B & I \\ I & B^{-1}\end{bmatrix}\succcurlyeq 0$ if $B\succ 0$.

11. Consider (7.7.7) when $A=C$. Show that the following are equivalent:

(a) $(x^*Ax)(y^*Ay)\geq |x^*By|^2$ for all $x, y\in \mathbf{C}^n$
(b) $x^*Ax+y^*Ay\geq \frac{1}{2}|x^*By|^2$ for all $x, y\in \mathbf{C}^n$
(c) $\rho(B^*A^{-1}BA^{-1})\leq 1$
(d) $x^*Ax\geq |x^*Bx|$ for all $x\in \mathbf{C}^n$

12. Show that if $A\in M_n$ is invertible and symmetric, then all the row sums of $A^{-1}\circ A$ are equal to 1. *Hint:* Look at the cofactor representation for the elements of A^{-1}. Thus, if A is real and positive definite, show that it is not possible to have $A^{-1}\circ A\succ I$ even though $A^{-1}\circ A\succcurlyeq I$.

13. If $A^{(k)}$ denotes the Hadamard kth power of A, and if $A\in M_n$ is positive definite, show that $(A^{-1})^{(k)}\succcurlyeq (A^{(k)})^{-1}$ for all $k=1, 2, \ldots$.

Further Reading. For the background to (7.7.7) and further references see C. FitzGerald and R. Horn, "On the Structure of Hermitian-Symmetric Inequalities," *J. London Math. Soc.* 15(2) (1977), 419–430. See also C. Johnson, "Partitioned and Hadamard Product Matrix Inequalities," *J. Research NBS* 83 (1978), 585–591, for further references related to (7.7.8), (7.7.9).

7.8 Inequalities for positive definite matrices

We next discuss inequalities involving quantities associated with one or more positive definite matrices. These are to be distinguished from the matrix inequalities introduced in the preceding section, although examples of the former are often associated with instances of the latter. For example, $A\succcurlyeq B\succcurlyeq 0$ implies $\det A\geq \det B$. The positive definite matrices are rich in inequalities involving determinants, eigenvalues, and other quantities. In this section, we examine some inequalities that do not necessarily stem from matrix inequalities.

The fundamental determinant inequality for positive definite matrices is Hadamard's inequality. Many other inequalities are generalizations in one way or another of this one result.

7.8.1 Theorem (Hadamard's inequality). If $A=[a_{ij}]\in M_n$ is positive semidefinite, then

$$\det A\leq \prod_{i=1}^{n} a_{ii}$$

Furthermore, when A is positive definite, then equality holds if and only if A is diagonal.

Proof: If A is singular, there is nothing to prove, so assume A is nonsingular, in which case all $a_{ii}\neq 0$. Define $d_i\equiv a_{ii}^{-1/2}$ and let $D=\operatorname{diag}(d_1, d_2, \ldots, d_n)$. Since $\det DAD\leq 1$ if and only if $\det A\leq a_{11}a_{22}\cdots a_{nn}$, it suffices to assume that each diagonal entry of A is equal to 1. If $\lambda_1, \ldots, \lambda_n$ are the (necessarily positive) eigenvalues of A, we have

$$\det A=\prod_{i=1}^{n}\lambda_i\leq\left(\frac{1}{n}\sum_{i=1}^{n}\lambda_i\right)^n=\left(\frac{1}{n}\operatorname{tr} A\right)^n=1$$

The inequality follows from the arithmetic-geometric mean inequality for nonnegative real numbers. Equality holds in the arithmetic-geometric mean inequality if and only if all $\lambda_i=1$, but since A is Hermitian and hence diagonalizable, this can occur if and only if $A=I$. Thus, equality holds in the original inequality when A is positive definite if and only if A is diagonal. □

Another determinantal inequality for general square matrices is equivalent to (7.8.1), and it also is referred to as Hadamard's inequality. Geometrically, $|\det A|$ is the volume of the n-dimensional parallelepiped whose generating edges are given by the rows (or columns) of A. This volume is largest when the generating edges are orthogonal, and in this case the volume is the product of the lengths of the edges. Hadamard's inequality is an algebraic statement of this geometric inequality.

7.8.2 Corollary (Hadamard's inequality). For any matrix $B=[b_{ij}]\in M_n$,

$$|\det B|\leq \prod_{i=1}^{n}\left(\sum_{j=1}^{n}|b_{ij}|^2\right)^{1/2} \quad\text{and}\quad |\det B|\leq \prod_{j=1}^{n}\left(\sum_{i=1}^{n}|b_{ij}|^2\right)^{1/2}$$

Furthermore, when B is nonsingular, then equality holds if and only if the rows (respectively, columns) of B are orthogonal.

Proof: If B is singular, there is nothing to prove. If B is nonsingular, apply (7.8.1) to the positive definite matrix $A \equiv BB^*$, and take square roots. The right-hand side of the first inequality is the square root of the product of the diagonal entries of A, and the left-hand side is the square root of $\det A$. The rows of B are orthogonal exactly when A is diagonal, which is the case of equality in (7.8.1). The second inequality follows from applying the first to B^*. □

Exercise. We have deduced (7.8.2) from (7.8.1). Now show that (7.8.1) follows from (7.8.2). *Hint:* If A is positive definite, there exists a unique positive definite B such that $B^2 = A$. Apply (7.8.2) to B and square.

Exercise. Use Hadamard's inequalities (and variants thereof) to give the best bound you can for

$$\det \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & 1 \end{bmatrix}$$

Two generalizations that refine Hadamard's inequality for positive definite matrices are attributed to Fischer and to Szasz. In Fischer's inequality, complementary principal submatrices play the role that diagonal entries play in Hadamard's inequality.

7.8.3 **Theorem** (Fischer's inequality). Suppose that

$$P = \begin{bmatrix} A & B \\ B^* & C \end{bmatrix}$$

is a positive definite matrix that is partitioned so that A and C are square and nonempty. Then

$$\det P \leq (\det A)(\det C)$$

Proof: Let $X = -A^{-1}B$, and compute

$$\det P = \det \begin{bmatrix} I & 0 \\ X^* & I \end{bmatrix} \begin{bmatrix} A & B \\ B^* & C \end{bmatrix} \begin{bmatrix} I & X \\ 0 & I \end{bmatrix} = \det \begin{bmatrix} A & 0 \\ 0 & C - B^*A^{-1}B \end{bmatrix}$$

$$= (\det A)(\det[C - B^*A^{-1}B]) \leq (\det A)(\det C)$$

The last inequality utilizes (7.7.6) and (7.7.4b) to ensure that $\det C \geq \det(C - B^*A^{-1}B)$ because $C \succcurlyeq C - B^*A^{-1}B \succcurlyeq 0$.

Exercise. Deduce Hadamard's inequality (7.8.1) from Fischer's inequality. Also, formulate and state Fischer's inequality for partitions of P finer than that in (7.8.3) (two principal submatrices) but not so fine as in (7.8.1) (n principal submatrices). Note that in this case the right-hand side of Fischer's inequality is less than or equal to the right-hand side of Hadamard's inequality. Thus, Fischer's inequality for various partitions of increasing refinement gives a monotone nondecreasing sequence of upper bounds on $\det P$.

There is another inequality that gives a sequence of upper bounds on the determinant and includes Hadamard's bound. Let $P_k(A)$ denote the product of all the k-by-k principal minors of A [there are $\binom{n}{k}$ of them]. Notice that $P_n(A) = \det A$ and $P_1(A) = a_{11}a_{22} \cdots a_{nn}$.

7.8.4 **Theorem** (Szasz's inequality). If $A \in M_n$ is positive definite, then

$$P_{k+1}(A)^{\binom{n-1}{k}^{-1}} \leq P_k(A)^{\binom{n-1}{k-1}^{-1}} \quad \text{for all} \quad k = 1, 2, \ldots, n-1$$

Proof: Since the diagonal entries of A^{-1} are just ratios of $(n-1)$-by-$(n-1)$ principal minors of A to $\det A$, direct application of (7.8.1) to the positive definite matrix A^{-1} yields

$$\frac{1}{\det A} = \det A^{-1} \leq \frac{P_{n-1}(A)}{(\det A)^n}$$

and hence

$$P_n(A)^{n-1} = (\det A)^{n-1} \leq P_{n-1}(A)$$

Extraction of $(n-1)$st roots of both sides of this inequality produces the case $k = n-1$ of Szasz's family of inequalities. The remaining cases may be derived inductively. For example, for the case $k = n-2$, one identifies each $(n-1)$-by-$(n-1)$ principal submatrix as one initial matrix and applies the above inequality to get

$$P_{n-1}(A)^{n-2} \leq P_{n-2}(A)^2$$

since each $(n-2)$-by-$(n-2)$ principal submatrix of A occurs twice as a principal submatrix of some $(n-1)$-by-$(n-1)$ principal submatrix of A. Extraction of the $(n-1)(n-2)$st root of both sides yields the case $k = n-2$, and the remaining cases follow in the same way. □

Exercise. Show that Szasz's inequality implies Hadamard's inequality (7.8.1). What is the case of equality?

7.8.5 **Observation.** Let $A \in M_n$ be positive semidefinite and define

$$\alpha(A) \equiv \begin{cases} \dfrac{\det A}{\det A_{11}} & \text{if } A_{11} \text{ is positive definite} \\ 0 & \text{otherwise} \end{cases}$$

where A_{11} is the $(n-1)$-by-$(n-1)$ principal submatrix of A that results from deleting the first row and column of A. Let $E_{11} \in M_n$ be the matrix whose 1, 1 entry is 1 and all of whose remaining entries are 0. Then $A - tE_{11}$ is positive semidefinite for all $t \le \alpha(A)$ and is not positive semidefinite for any $t > \alpha(A)$; in particular, $A - \alpha(A)E_{11}$ is positive semidefinite.

Proof: It suffices to consider the case in which A is positive definite. Use (7.2.5) applied to the "trailing" principal minors. Notice that the first $n-1$ trailing minors of $A - tE_{11}$ are the same as those of A, and that $\det(A - tE_{11}) = \det A - t \det A_{11}$. □

Exercise. Provide details for the proof of (7.8.5).

Exercise. Prove Hadamard's inequality (7.8.1) by induction using (7.8.5).

Hadamard's inequality (7.8.1) may also be stated in terms of Hadamard products as

$$(\det A) \prod_{i=1}^{n} 1 \le \det A \circ I$$

The following is an inequality of Oppenheim (strengthened by Schur) that generalizes Hadamard's inequality by showing that there is nothing special about the role of the identity matrix in the above inequality.

7.8.6 **Theorem** (Oppenheim's inequality). If $A, B \in M_n$ are positive semi-definite, then

$$(\det A) \prod_{i=1}^{n} b_{ii} \le \det A \circ B$$

Proof: We proceed by induction on n. The assertion is immediate for $n = 1$. If $n \ge 2$ and if it holds for all matrices of size at most $n-1$, then it follows from the induction hypothesis that

$$(\det A_{11}) \prod_{i=2}^{n} b_{ii} \le \det A_{11} \circ B_{11}$$

where the notation is the same as in (7.8.5). Notice that $A_{11} \circ B_{11} = (A \circ B)_{11}$. Since $A - \alpha E_{11}$ is positive semidefinite, we know that $(A - \alpha E_{11}) \circ B$ is positive semidefinite, and hence

$$0 \le \det(A - \alpha E_{11}) \circ B = (\det A \circ B) - \alpha b_{11}(\det A_{11} \circ B_{11})$$

From this it follows that

$$\det A \circ B \ge \alpha b_{11} \det A_{11} \circ B_{11} \ge \alpha b_{11} (\det A_{11}) \prod_{i=2}^{n} b_{ii} = (\det A) \prod_{i=1}^{n} b_{ii} \qquad \square$$

Exercise. If $A, B \in M_n$ are positive definite, show that

$$(\det A)(\det B) \le \det A \circ B$$

and show further that

$$(\det A)(\det B) \le (\det A) \prod_{i=1}^{n} b_{ii} \le \det A \circ B \le \prod_{i=1}^{n} a_{ii} \prod_{i=1}^{n} b_{ii}$$

Exercise. If $A \in M_n$ is positive definite, show that $\det A \circ A^{-1} \ge 1$.

A determinantal inequality of a rather different sort applies to non-Hermitian matrices A for which $H(A)$ is positive definite. It may be thought of as a generalization of the inequality $|z| \ge |\operatorname{Re} z|$ for complex numbers.

7.8.7 **Theorem** (Ostrowski–Taussky). If $A \in M_n$ is such that $H(A) \equiv (A + A^*)/2$ is positive definite, then

$$\det H(A) \le |\det A|$$

Equality holds if and only if A is Hermitian.

Proof: Let $S(A) = (A - A^*)/2$, so that $A = H(A) + S(A)$. The asserted inequality is then the statement that

$$|\det[I + H(A)^{-1}S(A)]| \ge 1$$

But $H(A)^{-1}S(A)$ is similar to the skew-Hermitian matrix

$$H(A)^{-1/2}S(A)H(A)^{-1/2}$$

and hence it has only pure imaginary eigenvalues. Thus, it suffices to note that $|1 + it| \ge 1$ for any real number t. If $it_1, it_2, \ldots, it_n$ are the eigenvalues of $H(A)^{-1}S(A)$, then

$$|\det[I+H(A)^{-1}S(A)]| = \prod_{j=1}^{n} |1+it_j| \geq 1$$

Also, equality holds if and only if all $t_j=0$, which is equivalent to $S(A)=0$ since a skew-Hermitian matrix is diagonalizable. □

An important determinantal inequality involving the sum of two positive definite matrices is due to Minkowski. Its proof is similar to that of the preceding result.

7.8.8 **Theorem** (Minkowski's inequality). If $A, B \in M_n$ are positive definite, then

$$[\det(A+B)]^{1/n} \geq (\det A)^{1/n} + (\det B)^{1/n}$$

Proof: Observing that both sides of the asserted inequality are homogeneous of the same degree, we multiply on the right and left by $(\det A^{-1/2})^{1/n}$. Thus, we may assume without loss of generality that $A=I$, and we must prove that

$$[\det(I+B)]^{1/n} \geq 1 + (\det B)^{1/n}$$

If $0<\lambda_1 \leq \cdots \leq \lambda_n$ are the eigenvalues of B, the desired inequality is then equivalent to

$$\prod_{i=1}^{n} (1+\lambda_i) \geq (1+\sqrt[n]{\lambda_1 \cdots \lambda_n})^n$$

which may be verified directly by explicit multiplication of both sides and term-by-term comparison using the arithmetic–geometric mean inequality. □

Exercise. Provide details for the proof of (7.8.8) and show that equality holds if and only if $B=cA$ for some $c \geq 0$.

Problems

1. Inequality (7.8.2) says that the magnitude of the determinant is dominated by the product of the l_2 norms of the rows. Compare this with the result [see Problem 3 in Section (6.1)] that the magnitude of the determinant is dominated by the product of the l_1 norms of the rows. What does each bound say geometrically? Are there other such bounds? Try l_∞.

2. The left-hand side of (7.8.2) is invariant under left unitary multiplication of B, and the left-hand side of (7.8.1) is invariant under unitary similarities of A, but the right-hand sides are not so invariant. When are the right-hand sides minimized? When are they maximized? Can a better bound be obtained in this way?

3. Use Fischer's inequality to verify the following block generalization of Hadamard's inequality (7.8.2): Let $A=[A_{ij}]$ be an nk-by-nk complex matrix, partitioned so that each block $A_{ij} \in M_k$. Then

$$|\det A| \leq \left[\prod_{i=1}^{n}\left(\sum_{j=1}^{n} \|A_{ij}\|_2^2\right)^{1/2}\right]^k$$

May matrix norms other than the spectral norm be used here?

4. Determine the cases of equality in (7.8.6).

5. Let $A, B \in M_n$ be positive definite and show that

$$\det A \circ B + (\det A)(\det B) \geq (\det A)\prod_{i=1}^{n} b_{ii} + (\det B)\prod_{i=1}^{n} a_{ii}$$

Show that this strengthens (7.8.6). *Hint:* Show that

$$\alpha(A \circ B) \geq \alpha(A)b_{11} + \alpha(B)a_{11} - \alpha(A)\alpha(B)$$

and apply this to the natural induction hypothesis.

6. Show that the inequality in the previous problem can be extended further to yield

$$\det A \circ B + (\det A)(\det B)\frac{\det A_{11} \circ B_{11}}{(\det A_{11})(\det B_{11})}$$

$$\geq (\det A)\prod_{i=1}^{n} b_{ii} + (\det B)\prod_{i=1}^{n} a_{ii} + (\det A)b_{11}(\det B_{11})\left(\frac{a_{22}\cdots a_{nn}}{\det A_{11}} - 1\right)$$

$$+\left[\det(B)a_{11}(\det A_{11})\left(\frac{b_{22}\cdots b_{nn}}{\det B_{11}} - 1\right)\right]$$

7. If $A \in M_n$ has positive definite Hermitian part $H=(A+A^*)/2$ and if $n>1$, show that the inequality in (7.8.7) can be strengthened to

$$\det H(A) + |\det S(A)| \leq |\det A|$$

What is the case of equality? *Hint:* You must show that

$$|\det[I+H^{-1}(A)S(A)]| \geq 1 + |\det H(A)^{-1}S(A)|$$

which is equivalent to having

$$\prod_{j=1}^{n} |1+it_j| \geq 1 + \prod_{j=1}^{n} |t_j|$$

Show that

$$\prod_{j=1}^{n} |1+it_j|^2 = 1 + \sum_{j=1}^{n} t_j^2 + \cdots + \prod_{j=1}^{n} t_j^2$$

$$\geq 1 + n \prod_{j=1}^{n} t_j + \left(\prod_{j=1}^{n} t_j\right)^2 \geq \left(1 + \prod_{j=1}^{n} |t_j|\right)^2$$

Can you strengthen the inequality further? *Note:* A natural inequality for complex numbers of which the stated inequality may be thought of as a generalization would be $|z| \geq |\mathrm{Re}\, z| + |\mathrm{Im}\, z|$; show that this inequality is false (hence the assumption that $n > 1$) and conclude that the determinantal inequality is therefore somewhat surprising.

8. If $A, B \in M_n$ are positive definite, show that $\det(A+B) \geq \det A + \det B$.

9. Use Minkowski's inequality to prove Fischer's inequality. *Hint:* Apply Minkowski's inequality to the two positive definite matrices

$$\begin{bmatrix} A & B \\ B^* & C \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} \begin{bmatrix} A & B \\ B^* & C \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} = \begin{bmatrix} A & -B \\ -B^* & C \end{bmatrix}$$

10. A positive definite matrix $P \in M_n$ can be factored as $P = LL^*$, where L is lower triangular with positive diagonal entries (7.2.9). Use this fact to prove Fischer's inequality.

11. Let $A, B \in M_n$ be positive semidefinite. If A and B are nonsingular, show that $A \circ B$ is nonsingular (and positive definite). If $A \circ B$ is singular, show that at least one of A, B is singular. How does this relate to the inequality $\mathrm{rank}\, A \circ B \leq (\mathrm{rank}\, A)(\mathrm{rank}\, B)$ from Section (7.5)?

12. Show that if $A = [a_{ij}] \in M_3$ is a matrix with real entries and if all $|a_{ij}| \leq 1$, then $|\det A| \leq 3\sqrt{3}$. Also show that this bound is never attained. *Hint:*

$$\frac{\partial}{\partial a_{ij}} (\det A) = (-1)^{i+j} A_{ij} \quad \text{and} \quad \frac{\partial^2}{\partial a_{ij}^2} (\det A) \equiv 0$$

where A_{ij} is the determinant of A with row i and column j deleted. If $A_{ij} = 0$, then $\det A$ is independent of the value of a_{ij}, which may therefore be taken to be ± 1. If $A_{ij} \neq 0$, then $\det A$ will not have an extremum with respect to a_{ij} if $0 < a_{ij} < 1$. Thus, $|\det A|$ achieves its maximum value within the given constraints when all $a_{ij} = \pm 1$. There are only finitely many such matrices for $n = 3$. What is the result for general $n > 3$? If A has complex entries, then use the maximum principle (the maximum modulus theorem) for analytic functions to show that $|\det A|$ cannot have a maximum in the interior of the set $\{A \in M_n: \text{ all } |a_{ij}| \leq 1\}$.

13. If $A = [a_{ij}] \in M_n$ and $K = \max\{|a_{ij}|\}$, use Hadamard's inequality to show that $|\det A| \leq K^n n^{n/2}$.

14. Let $A \in M_n$ be positive definite and let $\alpha \subseteq N = \{1, \ldots, n\}$ be an index set. Fischer's inequality may be stated as $\det A \leq \det A(\alpha) \det A(\alpha')$, in which α' is the complement of α with respect to N. A generalization of this result, often referred to as the Hadamard–Fischer inequalities, is

$$\det A(\alpha \cup \beta) \leq \frac{\det A(\alpha) \det A(\beta)}{\det A(\alpha \cap \beta)} \tag{7.8.9}$$

which holds for positive definite Hermitian matrices A and all index sets $\alpha, \beta \subseteq N$. By convention, $\det A(\phi) \equiv 1$. Prove the Hadamard–Fischer inequalities using only Fischer's inequality and the second formula in (0.8.4). *Hint:* Assume without loss of generality that $\alpha \cup \beta = N$ and apply Fischer's inequality to $A^{-1}(\alpha' \cup \beta')$. Then apply (0.8.4) to each minor.

15. Use the fact that a positive definite Hermitian matrix may be written as LL^*, with L lower triangular and nonsingular (7.2.9), to give a direct proof of the Hadamard–Fischer inequalities (7.8.9). *Hint:* Suppose $1 \leq j < k < n$, and suppose that $\alpha = \{1, \ldots, k\}$ and $\beta = \{1, \ldots, j, k+1, \ldots, n\}$, without loss of generality. Then consider a corresponding block 3-by-3 partitioning of A and L.

16. Let $A \in M_n$ be positive definite. Use the Hadamard–Fischer inequalities (7.8.9) to show that

$$\det A \leq \frac{\prod_{i=1}^{n-1} \det A(\{i, i+1\})}{\prod_{i=2}^{n-1} a_{ii}}$$

17. Let $A \in M_n$ be positive definite. Show that

$$\det A = \min\left\{ \prod_{i=1}^{n} v_i^* A v_i : \{v_1, \ldots, v_n\} \subset \mathbf{C}^n \text{ is an orthonormal set} \right\}$$

Hint: Let $V = [v_1 \ldots v_n] \in M_n$ and apply (7.8.1) to $\tilde{A} \equiv V^*AV$.

18. Let $A \in M_n$ be positive definite and let $\{u_1, \ldots, u_n\} \subset \mathbf{C}^n$ be an orthonormal set. Use Problem 17 to show that $\{u_1, \ldots, u_n\}$ are eigenvectors of A and $\{u_1^*Au_1, \ldots, u_n^*Au_n\}$ are the corresponding eigenvalues of A if and only if

$$\det A = \prod_{i=1}^{n} u_i^* A u_i$$

19. If $A \in M_n$ is positive definite, show that

$$n(\det A)^{1/n} = \min\{\text{tr } AB : B \in M_n \text{ is positive definite and } \det B = 1\}$$

Hint: Write $A = U\Lambda U^*$, with $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$, all $\lambda_i > 0$, and unitary $U \in M_n$, so $\text{tr } AB = \text{tr } \Lambda(U^*BU)$. Then use the arithmetic-geometric mean inequality and Hadamard's inequality (7.8.1) to show that

$$\frac{1}{n}\sum_{i=1}^{n} \lambda_i b_{ii} \geq \left(\prod_{i=1}^{n} \lambda_i b_{ii}\right)^{1/n} = \left(\det A \prod_{i=1}^{n} b_{ii}\right)^{1/n} \geq [\det A]^{1/n}$$

with equality possible.

20. Use the quasi-linearization in Problem 19 to prove Minkowski's inequality (7.8.8).

21. Let $A \in M_n$ be positive semidefinite and write A in partitioned form as

$$A = \begin{bmatrix} a_{11} & x^* \\ x & \tilde{A} \end{bmatrix}$$

Use the reduction formula for the determinant in Problem 15 of Section (4.1) and Problem 11 in Section (7.2) to show that

$$\det A = a_{11} \det \tilde{A} - x^*(\text{adj } \tilde{A})x \leq a_{11} \det \tilde{A}$$

Use induction and this inequality to give another proof of Hadamard's inequality (7.8.1).

Further Reading. For more information on Theorem (7.8.7) see A. M. Ostrowski and O. Taussky, "On the Variation of the Determinant of a Positive Definite Matrix," *Proc. Kon. Nederl. Acad. Wetensch. Amsterdam,* Ser. A, 54 (1951), 383–385. A class of inequalities that relates $\det A$ to other generalized matrix functions (0.3.2) of A when A is positive definite and that also generalizes Hadamard's inequality (7.8.1) may be found in I. Schur, "Über endliche Gruppen und Hermitesche Formen," *Math. Z.* 1 (1918), 184–207.

CHAPTER 8

Nonnegative matrices

8.0 Introduction

Suppose there are $n \geq 2$ cities $C_1, \ldots, C_n$ among which migration takes place as follows: Simultaneously at 8:00 A.M. each day a constant fraction a_{ij} of the current population of city j moves to city i for all $i \neq j$; the fraction a_{jj} of the current population of city j remains in city j. Thus, if we denote the population of city i on day m by $p_i^{(m)}$, we have the recursive relation

$$p_i^{(m+1)} = a_{i1}p_1^{(m)} + \cdots + a_{in}p_n^{(m)}, \qquad i = 1, \ldots, n, \qquad m = 0, 1, \ldots$$

between the population distributions on days m and $m+1$. If we denote the n-by-n matrix of migration coefficients by $A = [a_{ij}]$ and the population distribution vector on day m by $p^{(m)} = [p_i^{(m)}]$, then

$$p^{(m+1)} = Ap^{(m)} = AAp^{(m-1)} = \cdots = A^{m+1}p^{(0)}, \qquad m = 0, 1, \ldots$$

where $p^{(0)}$ is the initial population distribution. Since the coefficients a_{ij} represent population fractions, we have $0 \leq a_{ij} \leq 1$ and $\sum_{i=1}^{n} a_{ij} = 1$ for each $j = 1, \ldots, n$.

In order to make sensible long-range plans for city services and capital investment, government officials wish to know how the total population $p = \sum_{i=1}^{n} p_i^{(0)}$ will be distributed far into the future; that is, they want to know about the asymptotic behavior of $p^{(m)}$ for large m. But since $p^{(m)} = A^m p^{(0)}$, it is apparent that one must look at the asymptotic behavior of A^m.

As an example, let us consider in detail the case $n = 2$. We have $a_{11} + a_{21} = 1 = a_{12} + a_{22}$, so if we denote $a_{21} = \alpha$ and $a_{12} = \beta$, we have

$$A=\begin{bmatrix}1-\alpha & \beta\\ \alpha & 1-\beta\end{bmatrix}$$

and we are interested in A^m for large m. If A were diagonalizable, we could compute A^m explicitly. Thus, we begin by computing the eigenvalues of A: $\lambda_2=1$ and $\lambda_1=1-\alpha-\beta$. Since $0\le\alpha,\beta\le 1$, we have $\lambda_2=1\ge|\lambda_1|=|1-\alpha-\beta|$, so $1=|\lambda_2|=\rho(A)$ and the spectral radius of A is an eigenvalue of A. Moreover, except in the trivial case $\alpha=\beta=0$ (in which case A is reducible), we see that $\lambda_2=\rho(A)$ is a simple eigenvalue of A.

If $\alpha+\beta\neq 0$, the respective eigenvectors are $x=[\beta,\alpha]^T$ (for $\lambda_2=1$) and $z=[1,-1]^T$ (for λ_1), so in this case A is diagonalizable and $A=S\Lambda S^{-1}$, where

$$\Lambda=\begin{bmatrix}1 & 0\\ 0 & 1-\alpha-\beta\end{bmatrix},\qquad S=\begin{bmatrix}\beta & 1\\ \alpha & -1\end{bmatrix},\quad\text{and}\quad S^{-1}=\frac{1}{\alpha+\beta}\begin{bmatrix}1 & 1\\ \alpha & -\beta\end{bmatrix}$$

Notice that the components of the eigenvector x are nonnegative and are positive if A is irreducible.

If α and β are not both 1, then $|\lambda_1|=|1-\alpha-\beta|<1$ and so $\lambda_1^m\to 0$ as $m\to\infty$. Thus, in this case we have

$$\lim_{m\to\infty}A^m=S\Big(\lim_{m\to\infty}\Lambda^m\Big)S^{-1}=S\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}S^{-1}=\frac{1}{\alpha+\beta}\begin{bmatrix}\beta & \beta\\ \alpha & \alpha\end{bmatrix}$$

and so the equilibrium population distribution is

$$\lim_{m\to\infty}p^{(m)}=\frac{1}{\alpha+\beta}\begin{bmatrix}\beta & \beta\\ \alpha & \alpha\end{bmatrix}\begin{bmatrix}p_1^{(0)}\\ p_2^{(0)}\end{bmatrix}=\frac{1}{\alpha+\beta}\begin{bmatrix}\beta\\ \alpha\end{bmatrix}$$

Notice that the equilibrium distribution is entirely independent of the initial distribution. The matrix A^m approaches a limit whose columns are proportional to the eigenvector x associated with the eigenvalue 1 (which is the spectral radius of A), and the limiting population distribution is proportional to this same eigenvector.

The two exceptional cases not treated above are easily analyzed individually. If $\alpha=\beta=0$, then $A=I$, $\lim_{m\to\infty}A^m=I$, and $\lim_{m\to\infty}p^{(m)}=p^{(0)}$, so the limiting distribution is not independent of the initial distribution.

If $\alpha=\beta=1$, then $A=\left[\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right]$, and the two cities exchange their entire populations on successive days. The powers of A do not approach a limit and neither does the population distribution if the initial population distribution is unequal. However, there is a sense in which an "average equilibrium" is attained, namely

$$\lim_{m\to\infty}\frac{1}{m}\sum_{k=1}^{m}A^k=\begin{bmatrix}.5 & .5\\ .5 & .5\end{bmatrix}\quad\text{and}\quad\lim_{m\to\infty}\frac{1}{m}\sum_{k=1}^{m}p^{(k)}=\frac{p_1^{(0)}+p_2^{(0)}}{2}\begin{bmatrix}1\\ 1\end{bmatrix}$$

In summary, we found in this example that $\rho(A)=1$ and:

1. The spectral radius $\rho(A)=1$ is itself an eigenvalue of A, and is not just the absolute value of an eigenvalue.
2. The eigenvector x associated with the eigenvalue $\rho(A)$ can be taken to have nonnegative components, which are positive if A is irreducible.
3. $\rho(A)$ is a simple eigenvalue of strictly largest modulus if all the entries of A are positive.
4. If all entries of A are positive, then $\lim_{m\to\infty}[A/\rho(A)]^m$ exists and is a rank 1 matrix, all of whose columns are proportional to the eigenvector x.
5. In all cases, $\lim_{m\to\infty}(1/m)\sum_{k=1}^{m}(A/\rho(A))^k$ exists.

These conclusions are in fact generally true for $n\ge 2$, but it is not possible to analyze the general case with the simple direct methods employed above. For example, A need not be diagonalizable when $n\ge 2$ even if all the entries of A are positive. New tools are required and will be developed in the rest of this chapter.

Problems

1. Show that the matrix $A=\left[\begin{smallmatrix}1&1\\0&1\end{smallmatrix}\right]$ has spectral radius 1, but that A^m is unbounded as $m\to\infty$.

2. Consider the matrix

$$A_\epsilon=\begin{bmatrix}\dfrac{1}{1+\epsilon} & \dfrac{1}{1+\epsilon}\\[2mm] \dfrac{\epsilon^2}{1+\epsilon} & \dfrac{1}{1+\epsilon}\end{bmatrix},\qquad \epsilon>0$$

(a) Show that $\lambda_2=1$ is a simple eigenvalue of A_ϵ, that $\rho(A)=\lambda_2=1$, and that $1>|\lambda_1|$. (b) Show that

$$x=\frac{1}{1+\epsilon}\begin{bmatrix}1\\ \epsilon\end{bmatrix}\quad\text{and}\quad y=\frac{1+\epsilon}{2\epsilon}\begin{bmatrix}\epsilon\\ 1\end{bmatrix}$$

are eigenvectors of A_ϵ and A_ϵ^T, respectively, corresponding to the eigenvalue $\lambda=1$. (c) Calculate A_ϵ^m explicitly, $m=1,2,\ldots$. (d) Show that

$$\lim_{m\to\infty}A_\epsilon^m=\tfrac{1}{2}\begin{bmatrix}1 & \epsilon^{-1}\\ \epsilon & 1\end{bmatrix}$$

(e) Calculate xy^T and comment. (f) What happens if $\epsilon \to 0$? *Hint:* Set $B_\epsilon = (1+\epsilon)A_\epsilon$ and proceed as in the text to diagonalize B_ϵ.

3. Interpret what it means for the general matrix of intercity migration coefficients to be irreducible in terms of the freedom of travel of the populace.

4. For the two-city example discussed in this section, show that $\lim_{m\to\infty}(1/m)\sum_{k=1}^{m} A^k$ exists in the cases $\alpha=\beta=0$ and $\alpha+\beta\neq 0$. What is the limit in each case?

Further Readings. For a wealth of information about properties of positive and nonnegative matrices as well as many references to the theoretical and applied literature, see [BPl] and [Sen]. In [Var] there is a summary of results about nonnegative matrices, with special emphasis on applications to numerical analysis.

8.1 Nonnegative matrices – inequalities and generalities

Let $B=[b_{ij}]\in M_{n,r}$ and $A=[a_{ij}]\in M_{n,r}$. We write

$$B\geq 0 \quad \text{if all} \quad b_{ij}\geq 0$$

$$B>0 \quad \text{if all} \quad b_{ij}>0$$

$$A\geq B \quad \text{if} \quad A-B\geq 0$$

$$A>B \quad \text{if} \quad A-B>0$$

The reverse relations $\leq$ and $<$ are defined similarly. We define $|A|\equiv [|a_{ij}|]$. If $A\geq 0$, we say A is a *nonnegative* matrix, and if $A>0$, we say that A is a *positive* matrix. The following simple facts follow immediately from the definitions.

Exercise. Let $A,B\in M_{n,r}$. Show that

(8.1.1) $|A|\geq 0$ for every A; $|A|=0$ if and only if $A=0$.
(8.1.2) $|aA|=|a|\,|A|$ for all $a\in \mathbf{C}$.
(8.1.3) $|A+B|\leq |A|+|B|$.
(8.1.4) If $A\geq 0$ and $A\neq 0$, then it need not be true that $A>0$ if either n or r is greater than 1.
(8.1.5) If $A\geq 0$, $B\geq 0$, and $a,b\geq 0$, then $aA+bB\geq 0$.
(8.1.6) If $A\geq B$ and $C\geq D$, then $A+C\geq B+D$.
(8.1.7) If $A\geq B$ and $B\geq C$, then $A\geq C$.

Exercise. Now assume that $A,B,C,D\in M_n$ and that $x,y\in \mathbf{C}^n$. Show that:

(8.1.8) $|Ax|\leq |A|\,|x|$.
(8.1.9) $|AB|\leq |A|\,|B|$.
(8.1.10) $|A^m|\leq |A|^m$ for all $m=1,2,\ldots$.
(8.1.11) If $0\leq A\leq B$ and $0\leq C\leq D$, then $0\leq AC\leq BD$.
(8.1.12) If $0\leq A\leq B$, then $0\leq A^m\leq B^m$ for all $m=1,2,\ldots$.
(8.1.13) If $A\geq 0$, then $A^m\geq 0$; if $A>0$, then $A^m>0$ for all $m=1,2,\ldots$.
(8.1.14) If $A>0$, $x\geq 0$, and $x\neq 0$, then $Ax>0$.
(8.1.15) If $A\geq 0$, $x>0$, and $Ax=0$, then $A=0$.
(8.1.16) If $|A|\leq |B|$, then $\|A\|_2\leq \|B\|_2$.
(8.1.17) $\|A\|_2=\||A|\|_2$.

Obviously, the last two assertions hold for any absolute norm, of which the Frobenius norm (l_2 norm) is only one example. The first application of these simple relations is to an inequality for the spectral radius.

8.1.18 Theorem. Let $A,B\in M_n$. If $|A|\leq B$, then $\rho(A)\leq \rho(|A|)\leq \rho(B)$.

Proof: For every $m=1,2,\ldots$ we have $|A^m|\leq |A|^m\leq B^m$ by (8.1.10) and (8.1.12). Thus, by (8.1.16) and (8.1.17) we have

$$\|A^m\|_2\leq \||A|^m\|_2\leq \|B^m\|_2 \quad \text{and} \quad \|A^m\|_2^{1/m}\leq \||A|^m\|_2^{1/m}\leq \|B^m\|_2^{1/m}$$

for all $m=1,2,\ldots$. If we now let $m\to\infty$ and apply (5.6.14), we deduce that $\rho(A)\leq \rho(|A|)\leq \rho(B)$. □

8.1.19 Corollary. Let $A,B\in M_n$. If $0\leq A\leq B$, then $\rho(A)\leq \rho(B)$.

8.1.20 Corollary. Let $A\in M_n$. If $A\geq 0$ and if $\tilde{A}$ is any principal submatrix of A, then $\rho(\tilde{A})\leq \rho(A)$. In particular, $\max_{i=1,\ldots,n} a_{ii}\leq \rho(A)$.

Proof: Let $1\leq r\leq n$ and let $\tilde{A}$ be an r-by-r principal square submatrix of A. Let $\hat{A}$ denote the n-by-n matrix formed by placing the entries of $\tilde{A}$ in their former positions (as entries of A) and placing 0's elsewhere. Then $\rho(\tilde{A})=\rho(\hat{A})$ and $0\leq \hat{A}\leq A$, so $\rho(\tilde{A})=\rho(\hat{A})\leq \rho(A)$ by Corollary (8.1.19). □

The lower bound $a_{ii}\leq \rho(A)$ in the preceding corollary is the first nontrivial lower bound we have obtained on the spectral radius of a not-necessarily-Hermitian matrix, but the hypothesis that A is nonnegative is essential.

Exercise. Construct a matrix that is similar to $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and has no zero entries. What is its spectral radius? Is it nonnegative? What does this show about the last part of Corollary (8.1.20)?

Exercise. Show that if $A, B \in M_n$ and $0 \le A < B$, then $\rho(A) < \rho(B)$. *Hint:* There is some $\alpha > 1$ such that $0 \le A \le \alpha A < B$. The conclusion follows from Corollary (8.1.19) if $\rho(A) \ne 0$, and from Corollary (8.1.20) applied to B if $\rho(A) = 0$.

Since we shall soon have rather good upper bounds on the spectral radius of a nonnegative matrix, Theorem (8.1.18) will be useful in obtaining upper bounds on the spectral radius of an arbitrary matrix.

8.1.21 **Lemma.** Let $A \in M_n$ and suppose that $A \ge 0$. If the row sums of A are constant, then $\rho(A) = \|\|A\|\|_\infty$. If the column sums of A are constant, then $\rho(A) = \|\|A\|\|_1$.

Proof: We know that $\rho(A) \le \|\|A\|\|$ for any matrix norm $\|\|\cdot\|\|$, but if the row sums are constant, $x = [1, \ldots, 1]^T$ is an eigenvector with eigenvalue $\|\|A\|\|_\infty$ and so $\rho(A) = \|\|A\|\|_\infty$. The statement for column sums follows from applying the same argument to A^T. □

8.1.22 **Theorem.** Let $A \in M_n$ and suppose $A \ge 0$. Then

$$\min_{1 \le i \le n} \sum_{j=1}^{n} a_{ij} \le \rho(A) \le \max_{1 \le i \le n} \sum_{j=1}^{n} a_{ij} \tag{8.1.23}$$

and

$$\min_{1 \le j \le n} \sum_{i=1}^{n} a_{ij} \le \rho(A) \le \max_{1 \le j \le n} \sum_{i=1}^{n} a_{ij} \tag{8.1.24}$$

Proof: Let $\alpha = \min_{1 \le i \le n} \sum_{j=1}^{n} a_{ij}$ and construct a new matrix B with $A \ge B \ge 0$ and $\sum_{j=1}^{n} b_{ij} \equiv \alpha$ for all $i = 1, 2, \ldots$. For example, if $\alpha = 0$, we set $B = 0$, and if $\alpha > 0$, we could set $b_{ij} = \alpha a_{ij} (\sum_{j=1}^{n} a_{ij})^{-1}$. By Lemma (8.1.21), $\rho(B) = \alpha$, and $\rho(B) \le \rho(A)$ by Corollary (8.1.19). The upper bound is easily established in a similar fashion. The column sum bounds follow from applying the row sum bounds to A^T. □

Exercise. Prove the upper bound assertions in the preceding result.

8.1.25 **Corollary.** Let $A \in M_n$. If $A \ge 0$ and $\sum_{j=1}^{n} a_{ij} > 0$ for all $i = 1, 2, \ldots, n$, then $\rho(A) > 0$. In particular, $\rho(A) > 0$ if $A > 0$ or if A is irreducible and nonnegative.

Exercise. Show that an irreducible matrix cannot have a zero row or a zero column.

Since $\rho(S^{-1}AS) = \rho(A)$ whenever S is invertible, we may generalize the above theorem by introducing some free parameters. If $S = \mathrm{diag}(x_1, \ldots, x_n)$ and if all $x_i > 0$, then $S^{-1}AS \ge 0$ if $A \ge 0$. Applying Theorem (8.1.22) to $S^{-1}AS = [a_{ij} x_j x_i^{-1}]$, we obtain the following more general result.

8.1.26 **Theorem.** Let $A \in M_n$ and suppose $A \ge 0$. Then for any positive vector $x \in \mathbf{C}^n$ we have

$$\min_{1 \le i \le n} \frac{1}{x_i} \sum_{j=1}^{n} a_{ij} x_j \le \rho(A) \le \max_{1 \le i \le n} \frac{1}{x_i} \sum_{j=1}^{n} a_{ij} x_j \tag{8.1.27}$$

and

$$\min_{1 \le j \le n} x_j \sum_{i=1}^{n} \frac{a_{ij}}{x_i} \le \rho(A) \le \max_{1 \le j \le n} x_j \sum_{i=1}^{n} \frac{a_{ij}}{x_i} \tag{8.1.28}$$

8.1.29 **Corollary.** Let $A \in M_n$, let $x \in \mathbf{R}^n$, and suppose that $A \ge 0$ and $x > 0$. If $\alpha, \beta \ge 0$ are such that $\alpha x \le Ax \le \beta x$, then $\alpha \le \rho(A) \le \beta$. If $\alpha x < Ax$, then $\alpha < \rho(A)$; if $Ax < \beta x$, then $\rho(A) < \beta$.

Proof: If $\alpha x \le Ax$, then $\alpha \le \min_{1 \le i \le n} x_i^{-1} \sum_{j=1}^{n} a_{ij} x_j$. We conclude that $\alpha \le \rho(A)$ by the theorem. If $\alpha x < Ax$, then there is some $\alpha' > \alpha$ such that $\alpha' x \le Ax$. In this event, $\rho(A) \ge \alpha' > \alpha$, so $\rho(A) > \alpha$. The upper bounds are verified similarly. □

Exercise. Complete the proof of Corollary (8.1.29).

8.1.30 **Corollary.** Let $A \in M_n$ and suppose that A is nonnegative. If A has a positive eigenvector, then the corresponding eigenvalue is $\rho(A)$; that is, if $Ax = \lambda x$ and $x > 0$ and $A \ge 0$, then $\lambda = \rho(A)$.

Proof: If $x > 0$ and $Ax = \lambda x$, then $\lambda \ge 0$ and $\lambda x \le Ax \le \lambda x$. But then $\lambda \le \rho(A) \le \lambda$ by Corollary (8.1.29). □

8.1.31 **Corollary.** Let $A \in M_n$ and suppose that A is nonnegative. If A has a positive eigenvector, then

$$\rho(A) = \max_{x > 0} \min_{1 \le i \le n} \frac{1}{x_i} \sum_{j=1}^{n} a_{ij} x_j = \min_{x > 0} \max_{1 \le i \le n} \frac{1}{x_i} \sum_{j=1}^{n} a_{ij} x_j \tag{8.1.32}$$

Exercise. Prove the preceding result. Use the positive eigenvector x in (8.1.27).

8.1.33 **Corollary.** Let $A \in M_n$ and suppose that A is nonnegative. If A has a positive eigenvector x, then for all $m = 1, 2, \ldots$ and for all $i = 1, \ldots, n$ we have

$$\sum_{j=1}^{n} a_{ij}^{(m)} \le \left[\frac{\max_{1\le k\le n} x_k}{\min_{1\le k\le n} x_k}\right]\rho(A)^m \quad \text{and} \quad \left[\frac{\min_{1\le k\le n} x_k}{\max_{1\le k\le n} x_k}\right]\rho(A)^m \le \sum_{j=1}^{n} a_{ij}^{(m)} \tag{8.1.34}$$

where $A^m \equiv [a_{ij}^{(m)}]$. In particular, if $\rho(A) > 0$, the entries of $[\rho(A)^{-1}A]^m$ are uniformly bounded for $m = 1, 2, \ldots$.

Proof: If $Ax = \rho(A)x$, then $A^m x = \rho(A)^m x$. If $A \ge 0$, then $A^m \ge 0$ and we have

$$\rho(A)^m \max_{1\le k\le n} x_k \ge \rho(A^m)x_i = [A^m x]_i = \sum_{j=1}^{n} a_{ij}^{(m)} x_j$$

$$\ge \left(\min_{1\le k\le n} x_k\right) \sum_{j=1}^{n} a_{ij}^{(m)}$$

for any $i = 1, 2, \ldots, n$. Since $x > 0$, the asserted upper bound follows from division. Similarly, we have

$$\rho(A)^m \min_{1\le k\le n} x_k \le \rho(A)^m x_i = [A^m x]_i = \sum_{j=1}^{n} a_{ij}^{(m)} x_j$$

$$\le \left(\max_{1\le k\le n} x_k\right) \sum_{j=1}^{n} a_{ij}^{(m)}$$

for any $i = 1, \ldots, n$ and the asserted lower bound follows from division since $x > 0$. ☐

Problems

1. If $A \ge 0$ and if $A^k > 0$ for some k, show that $\rho(A) > 0$.

2. Give an example of a 2-by-2 matrix A such that $A \ge 0$, A is not positive, and $A^2 > 0$.

3. Suppose that $A \ge 0$ and $A \ne 0$. If A has a positive eigenvector, show that $\rho(A) > 0$.

4. If $\rho(A) < 1$, we know that $A^m \to 0$ as $m \to \infty$. If $A \ge 0$ and if A has a positive eigenvector, use Corollary (8.1.33) to show that $|A^m| \le \rho(A)^m C(A)$ for all $m = 1, 2, \ldots$, where $C(A)$ is a constant matrix. Show that the assumption that A has a positive eigenvector cannot be omitted by considering $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$. Discuss both of these results in the light of Corollary (5.6.13).

5. If $A \ge 0$ has a positive eigenvector, show that A is similar to a nonnegative matrix all of whose row sums are constant. What is this constant? *Hint:* Use the remarks preceding Theorem (8.1.26).

6. We shall show in Section (8.4) that a nonnegative irreducible matrix must have a positive eigenvector. Show that a nonnegative matrix can have a positive eigenvector and be reducible.

7. Let $A = [a_{ij}] \in M_n$ be nonnegative and have a positive eigenvector $x = [x_i]$. Use (8.1.33) to show that

(a) $$\frac{1}{n}\rho(A)^m \left[\frac{\min_{1\le k\le n} x_k}{\max_{1\le k\le n} x_k}\right] \le \max_{1\le p\le n} a_{ip}^{(m)} \quad \text{for each} \quad i = 1, 2, \ldots, n$$

(b) $$\lim_{m\to\infty} \left[\sum_{p=1}^{n} a_{ip}^{(m)}\right]^{1/m} = \rho(A) \quad \text{for each} \quad i = 1, 2, \ldots, n$$

We denote $A^m = [a_{ij}^{(m)}]$ for all $m = 1, 2, \ldots$.

8.2 Positive matrices

The theory of nonnegative matrices assumes its simplest and most elegant form for positive matrices, and it is for this case that O. Perron made the fundamental discoveries in 1907.

8.2.1 **Lemma.** Let $A \in M_n$ and suppose that $A > 0$, $Ax = \lambda x$, $x \ne 0$, and $|\lambda| = \rho(A)$. Then $A|x| = \rho(A)|x|$ and $|x| > 0$.

Proof: Compute

$$\rho(A)|x| = |\lambda|\,|x| = |\lambda x| = |Ax| \le |A|\,|x| = A|x|$$

so that $y \equiv A|x| - \rho(A)|x| \ge 0$. Since $|x| \ge 0$ and $|x| \ne 0$, we know from (8.1.14) that $A|x| > 0$. Corollary (8.1.25) also guarantees that $\rho(A) > 0$, so if $y = 0$, we have $A|x| = \rho(A)|x|$ and $|x| = \rho(A)^{-1}A|x| > 0$. If $y \ne 0$, set $z \equiv A|x| > 0$ and apply (8.1.14) again:

$$0 < Ay = Az - \rho(A)z \quad \text{or} \quad Az > \rho(A)z$$

But then by Corollary (8.1.29) we have the absurdity $\rho(A) > \rho(A)$. We conclude that $y = 0$ and we are done. ☐

From this technical result we easily deduce the first principal result about positive matrices.

8.2.2 Theorem. Let $A \in M_n$ and suppose that A is positive. Then $\rho(A) > 0$, $\rho(A)$ is an eigenvalue of A, and there is a positive vector x such that $Ax = \rho(A)x$.

Proof: There is an eigenvalue λ with $|\lambda| = \rho(A) > 0$ and an associated eigenvector $x \neq 0$. By the lemma, the required vector is $|x|$. □

Exercise. If $A \in M_n$ and $A > 0$, use Corollary (8.1.31) to show that

$$\rho(A) = \max_{x>0} \min_{1 \le i \le n} \frac{1}{x_i} \sum_{j=1}^{n} a_{ij} x_j = \min_{x>0} \max_{1 \le i \le n} \frac{1}{x_i} \sum_{j=1}^{n} a_{ij} x_j$$

By sharpening the statement of Lemma (8.2.1) slightly, we can improve our knowledge of the location of the eigenvalues of A.

8.2.3 Lemma. Let $A \in M_n$ and suppose that $A > 0$, $Ax = \lambda x$, $x \neq 0$, and $|\lambda| = \rho(A)$. Then for some $\theta \in \mathbf{R}$, $e^{-i\theta}x = |x| > 0$.

Proof: The hypothesis guarantees that $|Ax| = |\lambda x| = \rho(A)|x|$, and from Lemma (8.2.1) we know that $A|x| = \rho(A)|x|$ and $|x| > 0$. Combining these two identities and the triangle inequality, we have, for each $k = 1, \ldots, n$,

$$\rho(A)|x_k| = |\lambda|\,|x_k| = |\lambda x_k| = \left| \sum_{p=1}^{n} a_{kp} x_p \right|$$

$$\le \sum_{p=1}^{n} |a_{kp}|\,|x_p| = \sum_{p=1}^{n} a_{kp}|x_p| = \rho(A)|x_k|$$

Thus, equality must hold in the triangle inequality and hence the (nonzero) complex numbers $a_{kp}x_p$, $p = 1, \ldots, n$ must all lie on the same ray in the complex plane. If we denote their common argument by θ, then $e^{-i\theta}a_{kp}x_p > 0$ for all $p = 1, \ldots, n$. But since all $a_{kp} > 0$ we have $e^{-i\theta}x > 0$. □

8.2.4 Theorem. Let $A \in M_n$ and suppose A is positive. Then $|\lambda| < \rho(A)$ for every eigenvalue $\lambda \neq \rho(A)$.

Proof: By definition, $|\lambda| \le \rho(A)$ for all eigenvalues λ of A. Suppose $|\lambda| = \rho(A)$ and $Ax = \lambda x$, $x \neq 0$. By Lemma (8.2.3), $w \equiv e^{-i\theta}x > 0$ for some $\theta \in \mathbf{R}$, so $Aw = \lambda w$. But then $\lambda = \rho(A)$ by Corollary (8.1.30). □

We now know that if $A > 0$, then $\rho(A)$ is characterized as the eigenvalue of strictly largest modulus; there are no others. The next result says that $\rho(A)$ is an eigenvalue of geometric multiplicity 1; that is, the eigenspace corresponding to $\rho(A)$ has dimension 1. In fact, we shall see shortly that the algebraic multiplicity is also 1.

8.2.5 Theorem. Let $A \in M_n$ and suppose that $A > 0$ and that w and z are nonzero vectors such that $Aw = \rho(A)w$ and $Az = \rho(A)z$. Then there exists some $\alpha \in \mathbf{C}$ such that $w = \alpha z$.

Proof: By Lemma (8.2.3) there exist real numbers θ_1 and θ_2 such that $p \equiv e^{-i\theta_1}z > 0$ and $q \equiv e^{-i\theta_2}w > 0$. Set $\beta \equiv \min_{1\le i\le n} q_i p_i^{-1}$ and define $r \equiv q - \beta p$. Notice that $r \ge 0$ and at least one coordinate of r is 0, so r is not a positive vector. But $Ar = Aq - \beta Ap = \rho(A)q - \beta\rho(A)p = \rho(A)r$, so if $r \neq 0$, we know by (8.1.14) that $r = \rho(A)^{-1}Ar > 0$. Since this is not true, we conclude that $r = 0$ and hence $q = \beta p$ and $w = \beta e^{i(\theta_2 - \theta_1)}z$. □

8.2.6 Corollary. Let $A \in M_n$ and suppose that $A > 0$. There exists a unique vector x such that $Ax = \rho(A)x$, $x > 0$, and $\sum_{i=1}^{n} x_i = 1$.

Exercise. Prove Corollary (8.2.6).

The unique normalized eigenvector characterized in Corollary (8.2.6) is often called the *Perron vector* of A; $\rho(A)$ is often called the *Perron root* of A. Of course, A^T is a positive matrix if A is, so all the above results apply to A^T as well. The Perron vector of A^T is called the *left Perron vector* of A.

Exercise. If $A \in M_n$ and $A > 0$, and if there is some $x \in \mathbf{C}^n$ such that $x \ge 0$, $x \neq 0$, and $Ax = \lambda x$, show that x is a multiple of the Perron vector of A and that $\lambda = \rho(A)$.

We shall be interested in studying the behavior of powers A^m as $m \to \infty$ because these powers occur in applications to numerical analysis and to the theory of Markov chains in probability. The next lemma isolates the requirements that are essential to the various limit theorems about nonnegative matrices. Notice that all the hypotheses are met if $A > 0$ and $\lambda = \rho(A)$.

8.2.7 Lemma. Let $A \in M_n$ be given, let $\lambda \in \mathbf{C}$ be given, and suppose x and y are vectors such that

(1) $Ax = \lambda x$;
(2) $A^T y = \lambda y$; and
(3) $x^T y = 1$.

Define $L \equiv xy^T$. Then

(a) $Lx = x$ and $y^T L = y^T$;
(b) $L^m = L$ for all $m = 1, 2, \ldots$;
(c) $A^m L = LA^m = \lambda^m L$ for all $m = 1, 2, \ldots$;
(d) $L(A - \lambda L) = 0$;
(e) $(A - \lambda L)^m = A^m - \lambda^m L$ for all $m = 1, 2, \ldots$; and
(f) every nonzero eigenvalue of $A - \lambda L$ is also an eigenvalue of A.

If, in addition, we assume that

(4) $\lambda \neq 0$; and
(5) λ is an eigenvalue of A with geometric multiplicity 1;

then we also have that

(g) λ is not an eigenvalue of $A - \lambda L$; that is, $\lambda I - (A - \lambda L)$ is invertible.

Finally, if we assume that

(6) $|\lambda| = \rho(A) > 0$; and
(7) λ is the only eigenvalue of A with modulus $\rho(A)$;

and if we order the eigenvalues of A as $|\lambda_1| \leq |\lambda_2| \leq \cdots \leq |\lambda_{n-1}| < |\lambda_n| = |\lambda| = \rho(A)$, then

(h) $\rho(A - \lambda L) \leq |\lambda_{n-1}| < \rho(A)$;
(i) $(\lambda^{-1}A)^m = L + (\lambda^{-1}A - L)^m \to L$ as $m \to \infty$; and
(j) for every r such that $[|\lambda_{n-1}|/\rho(A)] < r < 1$ there exists some $C = C(r, A)$ such that $\|(\lambda^{-1}A)^m - L\|_\infty < Cr^m$ for all $m = 1, 2, \ldots$.

Proof: Conclusions (a), (b), and (c) follow directly from assumptions (1), (2), and (3); notice that (3) implies that both x and y are nonzero vectors. Conclusion (d) follows from (b) and (c). Statement (e) may be proved inductively using (b) and (c). If $\mu \neq 0$ is an eigenvalue of $A - \lambda L$ and if $(A - \lambda L)w = \mu w$ for some $w \neq 0$, then $L(A - \lambda L)w = 0w = 0 = \mu Lw$ and hence $Lw = 0$. Thus, $(A - \lambda L)w = Aw = \mu w$, so μ is also an eigenvalue of A and (f) is proved.

If we now invoke hypothesis (4) and take $\mu = \lambda$, this argument shows that if w were a λ eigenvector of $A - \lambda L$, then it would also be a λ eigenvector of A. On the basis of hypothesis (5) we must then conclude that $w = \alpha x$ for some $\alpha \neq 0$. But then $\mu w = \lambda w = (A - \lambda L)w = (A - \lambda L)\alpha x = \alpha\lambda x - \lambda\alpha x = 0$, which is impossible since $\lambda \neq 0$ and $w \neq 0$. This contradiction establishes (g). Because of (f), we know either that $\rho(A - \lambda L) = |\lambda_k|$ for some eigenvalue λ_k of A or that $\rho(A - \lambda L) = 0$. Since we have ordered the eigenvalues of A by increasing modulus and $|\lambda_n| = |\lambda| = \rho(A)$, we know in either event from (g) that $\rho(A - \lambda L) \leq |\lambda_{n-1}|$. Thus the inequality in (h) follows directly from (7). Combining (h) with (e), we calculate easily that $(\lambda^{-1}A - L)^m = (\lambda^{-1}A)^m - L \to 0$ as $m \to \infty$ since $\rho(\lambda^{-1}A - L) = \rho(A - \lambda L)/\rho(A) \leq |\lambda_{n-1}|/\rho(A) < 1$. The convergence rate in (j) is a direct consequence of Corollary (5.6.13) applied to the matrix $\lambda^{-1}A - L$ with ϵ chosen so that $\rho(\lambda^{-1}A - L) + \epsilon \leq [|\lambda_{n-1}|/\rho(A)] + \epsilon < r < 1$. □

Exercise. Supply the details in the proof of (a), (b), and (c) in the lemma.

8.2.8 **Theorem.** Let $A \in M_n$ and suppose that $A > 0$. Then

$$\lim_{m \to \infty} [\rho(A)^{-1}A]^m = L$$

where $L \equiv xy^T$, $Ax = \rho(A)x$, $A^T y = \rho(A)y$, $x > 0$, $y > 0$, and $x^T y = 1$.

Proof: The assumptions (1)–(7) of the lemma are met with $\lambda = \rho(A)$, x the Perron vector of A, and $y = (x^T z)^{-1} z$, where z is the Perron vector of A^T. The conclusion follows from (i). □

8.2.9 **Corollary.** If $A \in M_n$ and $A > 0$, then $L \equiv \lim_{m \to \infty} [\rho(A)^{-1}A]^m$ is a positive matrix of rank 1.

8.2.10 **Theorem.** If $A \in M_n$ and $A > 0$, then $\rho(A)$ is an eigenvalue of algebraic multiplicity 1; that is, $\rho(A)$ is a simple root of the characteristic equation $p_A(t) = 0$.

Proof: By the Schur triangularization theorem (2.3.1) we may write $A = U\Delta U^*$, where U is unitary, Δ is an upper triangular matrix with main diagonal entries $\rho, \ldots, \rho, \lambda_{k+1}, \ldots, \lambda_n$, and $\rho = \rho(A)$ is an eigenvalue of algebraic multiplicity $k \geq 1$; the eigenvalues λ_i all have modulus strictly less than $\rho(A)$ for all $i = k+1, \ldots, n$. But then

$$L = \lim_{m\to\infty} [\rho(A)^{-1}A]^m$$

$$= U \lim_{m\to\infty} \begin{bmatrix} 1 & & & & & \\ & \ddots & & & * & \\ & & 1 & & & \\ & & & \frac{\lambda_{k+1}}{\rho} & & \\ & 0 & & & \ddots & \\ & & & & & \frac{\lambda_n}{\rho} \end{bmatrix}^m U^*$$

$$= U \begin{bmatrix} 1 & & & & & \\ & \ddots & & * & & \\ & & 1 & & & \\ & & & 0 & & \\ 0 & & & & \ddots & \\ & & & & & 0 \end{bmatrix} U^*$$

where the diagonal entry 1 is repeated k times in the last two expressions and the diagonal entry 0 is repeated $n-k$ times in the last expression. Since the upper triangular matrix in the last expression has rank at least k, and since L has rank 1, we conclude that $k>1$ is impossible. □

We now summarize the principal results obtained in this section for positive matrices.

8.2.11 **Perron's theorem.** If $A \in M_n$ and $A>0$, then

(a) $\rho(A)>0$;
(b) $\rho(A)$ is an eigenvalue of A;
(c) There is an $x \in \mathbf{C}^n$ with $x>0$ and $Ax=\rho(A)x$;
(d) $\rho(A)$ is an algebraically (and hence geometrically) simple eigenvalue of A;
(e) $|\lambda|<\rho(A)$ for every eigenvalue $\lambda \neq \rho(A)$, that is, $\rho(A)$ is the unique eigenvalue of maximum modulus; and
(f) $[\rho(A)^{-1}A]^m \to L$ as $m\to\infty$, where $L \equiv xy^T$, $Ax=\rho(A)x$, $A^Ty=\rho(A)y$, $x>0$, $y>0$, and $x^Ty=1$.

Perron's theorem has many applications. One elegant and useful application is to obtain an eigenvalue inclusion region for a matrix A in terms of the spectral radius and main diagonal entries of a dominating nonnegative matrix.

8.2.12 **Theorem** (Ky Fan). Let $A=[a_{ij}]\in M_n$ and suppose that $B=[b_{ij}]\in M_n$ has nonnegative entries and $B \geq |A|$. Then every eigenvalue of A lies in the region

$$\bigcup_{i=1}^{n} \{z \in \mathbf{C}: |z-a_{ii}| \leq \rho(B)-b_{ii}\}$$

Proof: We may assume that $B>0$, for if some entries of B are zero we may consider $B_\epsilon \equiv [b_{ij}+\epsilon]$ for $\epsilon>0$; $B_\epsilon > |A|$, and $\rho(B_\epsilon)-(b_{ii}+\epsilon) \to \rho(B)-b_{ii}$ as $\epsilon \to 0$. By Perron's theorem there is a positive vector x such that $Bx=\rho(B)x$, and hence

$$\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|x_j \leq \sum_{\substack{j=1\\ j\neq i}}^{n} b_{ij}x_j = \rho(B)x_i - b_{ii}x_i \quad \text{for all} \quad i=1,2,\ldots,n$$

Thus, we have

$$\frac{1}{x_i}\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|x_j \leq \rho(B)-b_{ii} \quad \text{for all} \quad i=1,2,\ldots,n$$

The result follows from Corollary (6.1.6) with $p_i=x_i$. □

Part (f) of (8.2.11) guarantees that a certain limit exists, and (j) of (8.2.7) gives an upper bound on the rate of convergence

$$\|[\rho(A)^{-1}A]^m - L\|_\infty < Cr^m$$

for some positive constant C, which depends on A and r, for any r such that

$$\frac{|\lambda_{n-1}|}{\rho(A)} < r < 1$$

where λ_{n-1} is a second largest modulus eigenvalue of A. Even if $\rho(A)$ is known or easily estimated, it can be inconvenient or impossible to compute or estimate $|\lambda_{n-1}|$ in order to get a useful bound on the ratio $|\lambda_{n-1}|/\rho(A)$. In such a situation it can be useful to know an easily computed bound due to E. Hopf, which holds for any positive matrix $A=[a_{ij}]\in M_n$:

$$\frac{|\lambda_{n-1}|}{\rho(A)} \leq \frac{M-\mu}{M+\mu} < 1$$

where $M=\max\{a_{ij}: i,j=1,2,\ldots,n\}$ and $\mu=\min\{a_{ij}: i,j=1,2,\ldots,n\}$.

Problems

1. If $A>0$, if x is the Perron vector of A, and if z is the Perron vector of A^T, show that $x^Tz>0$.

2. If $\Delta \in M_n$ is an upper triangular matrix with k nonzero main diagonal entries, show that rank $\Delta \geq k$. Show by example that it is possible to have the rank greater than k under these conditions.

3. Apply the results derived in this section to the matrix

$$A = \begin{bmatrix} 1-\alpha & \beta \\ \alpha & 1-\beta \end{bmatrix}, \qquad 0 < \alpha, \beta < 1$$

and compare with the conclusions reached in Section (8.0).

4. Consider the general intercity migration model with $n > 2$ cities as described in Section (8.0). If all the migration coefficients a_{ij} are positive, what is the asymptotic behavior of the population distribution $p^{(m)}$ as $m \to \infty$?

5. If $A > 0$, describe in detail the asymptotic behavior of A^m as $m \to \infty$. *Hint:* There are three cases: $A^m \to 0$, A^m diverges, and A^m converges to a positive matrix. Characterize and analyze each case.

6. After Corollary (6.1.8) there was an exercise dealing with a 2-by-2 positive matrix. Discuss this example in light of the exercise following Theorem (8.2.2).

7. Let $A, B \in M_n$ and suppose that $A > B > 0$. Use the "min max" characterization of $\rho(B)$ to show that $\rho(A) > \rho(B)$. *Hint:* Let x be the Perron vector of A so that $Ax > Bx$.

8. If $A > 0$ and if x is the Perron vector of A, show that

$$\rho(A) = \sum_{i,j=1}^{n} a_{ij} x_j$$

Recall that $x_1 + \cdots + x_n = 1$ by definition.

9. If a positive matrix is nonsingular, show that the inverse matrix cannot be nonnegative. If a nonnegative matrix A is nonsingular, show that the inverse matrix can be nonnegative only if A has exactly one nonzero entry in each column. How is such a matrix related to a permutation matrix?

10. Provide the details for the following alternate proof of Theorem (8.2.10): If $\rho = \rho(A)$ has algebraic multiplicity $k > 1$ and if y and x are the left and right Perron vectors of A, respectively, then there is a nonzero vector z such that $x = (A - \rho I)z$. But then $y^T x = y^T (A - \rho I) z = 0^T z = 0$, which is impossible since $y^T x > 0$. *Hint:* Since the geometric multiplicity of ρ is 1, the Jordan form of $A - \rho I$ must have exactly one nilpotent block, which must have size at least 2. Show that any nilpotent matrix $B \in M_k$, $k > 1$, of rank $k-1$ has the property that if $Bu = 0$ for some $u \in \mathbf{C}^k$, then there is some $v \in \mathbf{C}^k$ such that $Bv = u$.

Further Readings. For a unified treatment of a family of bounds that includes Hopf's bound mentioned in the last paragraph of this section (and an extensive bibliography) see U. Rothblum and C. Tan, "Upper Bounds on the Maximum Modulus of Subdominant Eigenvalues of Nonnegative Matrices," *Linear Algebra Appl.* 66 (1985), 45–86. See also [Kel] chapter II, theorem 2.

8.3 Nonnegative matrices

Since one is confronted in practice with nonnegative matrices that are not positive, it is necessary to consider the extension of the theory developed in the preceding section to the case in which not all of the matrix entries are strictly positive. One might hope that this extension could be done by taking suitable limits, and this is the case for some results. But unfortunately, quantities such as rank and dimension are not continuous functions and so limit arguments are only partially applicable. The only results in Perron's theorem which generalize by taking limits are contained in the following theorem.

8.3.1 **Theorem.** If $A \in M_n$ and $A \geq 0$, then $\rho(A)$ is an eigenvalue of A and there is a nonnegative vector $x \geq 0$, $x \neq 0$, such that $Ax = \rho(A)x$.

Proof: For any $\epsilon > 0$, define $A(\epsilon) \equiv [a_{ij} + \epsilon] > 0$. Denote by $x(\epsilon)$ the Perron vector of $A(\epsilon)$, so $x(\epsilon) > 0$ and $\sum_{i=1}^{n} x(\epsilon)_i = 1$. Since the set of vectors $\{x(\epsilon) : \epsilon > 0\}$ is contained in the compact set $\{x : x \in \mathbf{C}^n, \|x\|_1 \leq 1\}$, there is a monotone decreasing sequence $\epsilon_1, \epsilon_2, \ldots$ with $\lim_{k \to \infty} \epsilon_k = 0$ such that $\lim_{k \to \infty} x(\epsilon_k) \equiv x$ exists. Since $x(\epsilon_k) > 0$ for all $k = 1, 2, \ldots$, it must be that $x = \lim_{k \to \infty} x(\epsilon_k) \geq 0$; $x = 0$ is impossible because

$$\sum_{i=1}^{n} x_i = \lim_{k \to \infty} \sum_{i=1}^{n} x(\epsilon_k)_i = 1$$

By Theorem (8.1.18), $\rho(A(\epsilon_k)) \geq \rho(A(\epsilon_{k+1})) \geq \cdots \geq \rho(A)$ for all $k = 1, 2, \ldots$, so the sequence of real numbers $\{\rho(A(\epsilon_k))\}_{k=1,2,\ldots}$ is a monotone decreasing sequence. Thus, $\rho \equiv \lim_{k \to \infty} \rho(A(\epsilon_k))$ exists and $\rho \geq \rho(A)$. But from the fact that

$$\begin{aligned} Ax &= \lim_{k \to \infty} A(\epsilon_k) x(\epsilon_k) = \lim_{k \to \infty} \rho(A(\epsilon_k)) x(\epsilon_k) \\ &= \lim_{k \to \infty} \rho(A(\epsilon_k)) \lim_{k \to \infty} x(\epsilon_k) = \rho x \end{aligned}$$

and the fact that $x \neq 0$, we deduce that ρ is an eigenvalue of A. But then $\rho \leq \rho(A)$, so it must be that $\rho = \rho(A)$. □

There is a generalization of part of the variational characterization (8.1.31) of the spectral radius to general nonnegative matrices and nonnegative vectors, but the proof is quite different.

8.3.2 **Theorem.** Let $A \in M_n$, $A \geq 0$, $x \in \mathbf{C}^n$, $x \geq 0$, and $x \neq 0$. If $Ax \geq \alpha x$ for some $\alpha \in \mathbf{R}$, then $\rho(A) \geq \alpha$.

Proof: Let $A = [a_{ij}]$, let $\epsilon > 0$, and define $A(\epsilon) \equiv [a_{ij} + \epsilon]$. Then $A(\epsilon) > 0$, so $A(\epsilon)$ has a positive left Perron vector $y(\epsilon)$; that is, $y(\epsilon)^T A(\epsilon) = \rho(A(\epsilon))y(\epsilon)^T$. We are given that $Ax - \alpha x \geq 0$, so $A(\epsilon)x - \alpha x > Ax - \alpha x \geq 0$ and hence $y(\epsilon)^T[A(\epsilon)x - \alpha x] = [\rho(A(\epsilon)) - \alpha]y(\epsilon)^T x \geq 0$. Since $y(\epsilon)^T x > 0$, we have $\rho(A(\epsilon)) - \alpha \geq 0$ for all $\epsilon > 0$. But $\rho(A(\epsilon)) \to \rho(A)$ as $\epsilon \to 0$, so we conclude that $\rho(A) \geq \alpha$. □

8.3.3 **Corollary.** If $A \in M_n$ and $A \geq 0$, then

$$\rho(A) = \max_{\substack{x \geq 0 \\ x \neq 0}} \min_{\substack{1 \leq i \leq n \\ x_i \neq 0}} \frac{1}{x_i} \sum_{j=1}^{n} a_{ij} x_j$$

Proof: If $A \geq 0$, $x \geq 0$, and $x \neq 0$, and if we choose

$$\alpha \equiv \min_{x_i \neq 0} \sum_{j=1}^{n} \frac{a_{ij} x_j}{x_i}$$

then $Ax \geq \alpha x$ and so $\alpha \leq \rho(A)$ by the theorem. But if we choose x to be the eigenvector whose existence is guaranteed by Theorem (8.3.1), then we see that this upper bound can be attained with $\alpha = \rho(A)$. □

Exercise. Consider $A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ and $x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ to show that the upper bound of (8.1.29) need not hold if x is not a positive vector. Show that the "min max" characterization in (8.1.32) is also false in general. As the preceding result shows, however, the "max min" characterization does generalize.

With an additional assumption, Theorem (8.3.2) can be strengthened slightly to give some information about the vector x.

8.3.4 **Theorem.** Let $A \in M_n$ and $A \geq 0$, and suppose that A has a positive left eigenvector. If $x \geq 0$, if $x \neq 0$, and if $Ax \geq \rho(A)x$, then $Ax = \rho(A)x$.

Proof: Let $y > 0$ be such that $A^T y = \rho(A)y$ and suppose that $x \geq 0$ is such that $x \neq 0$ and $Ax - \rho(A)x \geq 0$. Then

$$y^T[Ax - \rho(A)x] = \rho(A)y^T x - \rho(A)y^T x = 0$$

so it must be that $Ax - \rho(A)x = 0$. □

Without additional assumptions, we can go no further than Theorem (8.3.1) in generalizing Perron's theorem (8.2.11) to nonnegative matrices.

When $A \in M_n$ and $A \geq 0$, the nonnegative eigenvalue $\rho(A)$ is called the *Perron root* of A. Because an eigenvector associated with the Perron root of a nonnegative matrix is not necessarily uniquely determined, there is (unlike the situation when A is positive) no well-determined notion of "the Perron vector" for a general nonnegative matrix. For example, the nonnegative matrix $A = I$ has every nonnegative vector as an eigenvector associated with the Perron root $\rho(A) = 1$.

Problems

1. Show by example that the items from Perron's theorem (8.2.11) that are not included in Theorem (8.3.1) are not generally true of all nonnegative matrices. *Hint:* Consider $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$.

2. If $A \geq 0$ and $A^k > 0$ for some $k \geq 1$, show that A has a positive eigenvector.

3. If $A \geq 0$ has a nonnegative eigenvector x with $r \geq 1$ positive components and $n - r$ zero components, show that by a permutation similarity, A can be brought into the form $\begin{bmatrix} B & C \\ 0 & D \end{bmatrix}$, where $B \in M_r$, $C \in M_{r,(n-r)}$, $D \in M_{n-r}$; B, C, and D are nonnegative; and B has a positive eigenvector. If $r < n$, conclude that A must be reducible.

4. Show by example that the following generalization of Corollary (8.1.30) is false: Let $A \geq 0$. If A has a nonnegative eigenvector $x \geq 0$, $x \neq 0$, then $Ax = \rho(A)x$.

5. Consider the matrix $A = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$ and the vector $x = [1, 2]^T$. Show that Theorem (8.3.4) is not correct if we assume only that $A \geq 0$. What are the left and right Perron vectors of A?

6. If $A \geq 0$, show that there exists a positive matrix B that commutes with A if and only if A has left and right eigenvectors, each of which is positive. *Hint:* Let $B = xy^T$ if x and y are positive right and left eigenvectors of A. Conversely, if $x \geq 0$ and $Ax = \rho(A)x$, consider $BAx = ABx = B\rho(A)x > 0$.

7. If $A=[a_{ij}]\in M_n$ is nonnegative and tridiagonal, show that all the eigenvalues of A are real. *Hint:* First show that if all sub- and super-diagonal entries are positive, then a positive diagonal matrix D may be found so that $D^{-1}AD$ is symmetric. Then show that 0 entries above or below the diagonal are inconsequential.

8. Let a nonnegative $A\in M_n$ be given. Show that either A is irreducible or there exists a permutation matrix P such that

$$P^TAP=\begin{bmatrix} A_1 & & * \\ & \ddots & \\ 0 & & A_k \end{bmatrix}$$

in which each A_i is either irreducible or is the 1-by-1 zero matrix, $i=1,\dots,k$. This is called the *irreducible normal form* of A. Note that $\sigma(A)=\bigcup_{i=1}^{k}\sigma(A_i)$ and that the irreducible normal form of A is not necessarily unique.

9. A matrix $A=[a_{ij}]\in M_n(\mathbf{R})$ all of whose off-diagonal entries a_{ij}, $i\neq j$, are nonnegative is said to be *essentially nonnegative*. Show that if A is essentially nonnegative, then there is some $\lambda>0$ such that $\lambda I+A\geq 0$. Use this observation and (8.3.1) to show that if $A\in M_n$ is essentially nonnegative, then A has a real eigenvalue $r(A)$ (often called the *dominant eigenvalue* of A) with the property that $r(A)\geq \operatorname{Re}\lambda_i$ for every eigenvalue λ_i of A. Show that $r(A)$ need not be the eigenvalue of A with largest modulus, but when $A\geq 0$, $r(A)=\rho(A)$. *Hint:* The eigenvalues of $\lambda I+A$ are $\lambda+\lambda_i$.

10. Theorem (8.1.18) says that if $A\in M_n$ is nonnegative, then $\rho(A+B)\geq\rho(A)$ whenever $B\in M_n$ is also nonnegative; this is a sort of monotonicity result for the spectral radius. If $A\in M_n$ is essentially nonnegative (see Problem 9), show that $A+D$ is essentially nonnegative for all diagonal matrices $D\in M_n(\mathbf{R})$. If A is a given essentially nonnegative matrix and D is allowed to vary in the class of real diagonal matrices, it is known that the dominant eigenvalue $r(A+D)$ is a convex function of the diagonal entries of D.

Further Readings. See C. Johnson, R. Kellogg, and A. Stephens, "Complex Eigenvalues of a Nonnegative Matrix with a Specified Graph II," *Lin. Multilin. Alg.* 7 (1979), 129–143, and C. Johnson, "Row Stochastic Matrices Similar to Doubly Stochastic Matrices," *Lin. Multilin. Alg.* 10 (1981), 113–130 for results on the important topic of eigenvalue possibilities of nonnegative matrices. These papers include references

to classical work of Dmitriev, Dynkin, and Karpelevich and to the nonnegative inverse eigenvalue problem. The latter problem (of characterizing the sets of complex numbers that can be spectra of nonnegative matrices) is unsolved. For more information about Problem 10, see J. Cohen, "Convexity of the Dominant Eigenvalue of an Essentially Nonnegative Matrix," *Proc. Amer. Math. Soc.* 81 (1981), 657–658. See also Problem 15 in Section (8.4).

8.4 Irreducible nonnegative matrices

It is a useful heuristic principle that if one can prove a result for matrices with no 0 entries, then the result often generalizes to irreducible matrices. We have had one instance of this principle in the extensions of the basic Geršgorin theorem in Chapter 6, and we shall now have another. The basic idea has already been proved in Theorem (6.2.24); we restate the relevant portion here.

8.4.1 **Lemma.** Let $A\in M_n$ and suppose $A\geq 0$. Then A is irreducible if and only if $(I+A)^{n-1}>0$.

Exercise. If $A\in M_n$, show that A is irreducible if and only if A^T is irreducible.

For our purposes, we also need the following simple results.

8.4.2 **Lemma.** Let $A\in M_n$ and let $\lambda_1,\dots,\lambda_n$ be the eigenvalues of A (including multiplicities). Then $\lambda_1+1,\dots,\lambda_n+1$ are the eigenvalues of $I+A$ and $\rho(I+A)\leq 1+\rho(A)$. If $A\geq 0$, then $\rho(I+A)=1+\rho(A)$.

Proof: If $\lambda\in\sigma(A)$ has multiplicity k, then λ is a root of the characteristic equation $p_A(t)=\det(tI-A)=0$ of multiplicity k. But then $\lambda+1$ is a root of $p_{A+I}(s)=\det[sI-(A+I)]=0$ of multiplicity k because $\det(tI-A)=\det[(t+1)I-(A+I)]$. Thus, $\lambda_1+1,\dots,\lambda_n+1$ are the eigenvalues of $A+I$. Therefore, $\rho(I+A)=\max_{1\leq i\leq n}|\lambda_i+1|\leq\max_{1\leq i\leq n}|\lambda_i|+1=1+\rho(A)$. However, by (8.3.1), $1+\rho(A)$ is an eigenvalue of $I+A$ when $A\geq 0$, so that $\rho(I+A)=1+\rho(A)$ in this case. $\square$

Exercise. Explain why the following argument in support of the first part of the preceding lemma is incomplete: If λ is an eigenvalue of A, then

there is some vector $x \neq 0$ such that $Ax = \lambda x$. But then $(A+I)x = (\lambda+1)x$, so $\lambda+1$ is an eigenvalue of $A+I$.

8.4.3 **Lemma.** If $A \in M_n$, $A \geq 0$, and $A^k > 0$ for some $k \geq 1$, then $\rho(A)$ is an algebraically simple eigenvalue of A.

Proof: If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A, then $\lambda_1^k, \ldots, \lambda_n^k$ are the eigenvalues of A^k. We know that $\rho(A)$ is an eigenvalue of A by Theorem (8.3.1), so if $\rho(A)$ were a multiple eigenvalue of A, then $\rho(A)^k = \rho(A^k)$ would be a multiple eigenvalue of A^k. But this is impossible since $\rho(A^k)$ is a simple eigenvalue of A^k by Theorem (8.2.10). □

Now we shall see how much of Perron's theorem generalizes to nonnegative irreducible matrices. The name of Frobenius is associated with generalizations of Perron's results about positive matrices to nonnegative matrices.

8.4.4 **Theorem.** Let $A \in M_n$ and suppose that A is irreducible and nonnegative. Then

(a) $\rho(A) > 0$;
(b) $\rho(A)$ is an eigenvalue of A;
(c) There is a positive vector x such that $Ax = \rho(A)x$; and
(d) $\rho(A)$ is an algebraically (and hence geometrically) simple eigenvalue of A.

Proof: Corollary (8.1.25) shows that (a) holds under conditions even weaker than irreducibility. Assertion (b) holds for all nonnegative matrices A by Theorem (8.3.1), which also guarantees that there is a nonnegative vector $x \neq 0$ such that $Ax = \rho(A)x$. But then $(I+A)^{n-1}x = [1+\rho(A)]^{n-1}x$, and since the matrix $(I+A)^{n-1}$ is positive by Lemma (8.4.1), we see that the vector $(I+A)^{n-1}x$ must be positive by (8.1.14). Thus, $x = [1+\rho(A)]^{1-n}(I+A)^{n-1}x > 0$. To prove (d) we apply Lemma (8.4.2) to show that if $\rho(A)$ is a multiple eigenvalue of A, then $1+\rho(A) = \rho(I+A)$ is a multiple eigenvalue of $I+A$. But $I+A \geq 0$ and $(I+A)^{n-1} > 0$ by Lemma (8.4.1), so $1+\rho(A)$ must be a simple eigenvalue of $I+A$ by Lemma (8.4.3). □

The theorem guarantees that the eigenspace of an irreducible nonnegative matrix associated with the Perron root is one-dimensional. For an irreducible nonnegative matrix, the unique positive eigenvector whose components sum to 1 is called the *Perron vector.*

Since an irreducible nonnegative matrix has a positive eigenvector, the results at the end of Section (8.1) apply to this class of matrices. Of particular importance is the variational characterization (8.1.32) of the spectral radius. Moreover, A^T is irreducible if and only if A is irreducible, so an irreducible nonnegative matrix also has a positive left eigenvector. Thus, Theorem (8.3.4) holds for nonnegative irreducible matrices. This fact is crucial in the following extension of Theorem (8.1.18).

8.4.5 **Theorem.** Let $A, B \in M_n$. Assume that A is nonnegative and irreducible, and assume that $A \geq |B|$. Then $\rho(A) \geq \rho(B)$. If $\rho(A) = \rho(B)$ and if $\lambda = e^{i\varphi}\rho(B)$ is an eigenvalue of B, then there exist $\theta_1, \ldots, \theta_n \in \mathbf{R}$ such that $B = e^{i\varphi}DAD^{-1}$, where $D = \operatorname{diag}(e^{i\theta_1}, \ldots, e^{i\theta_n})$.

Proof: From Theorem (8.1.18) we know already that if $A \geq |B|$, then $\rho(A) \geq \rho(B)$. If $\rho(A) = \rho(B)$, then there exists some $x \neq 0$ such that $Bx = \lambda x$ with $|\lambda| = \rho(B) = \rho(A)$, and so

$$\rho(A)|x| = |\lambda x| = |Bx| \leq |B|\,|x| \leq A|x|$$

Since A is irreducible, we conclude from Theorem (8.3.4) that $A|x| = \rho(A)|x|$ and hence $|Bx| = |B|\,|x| = A|x|$. Furthermore, (c) and (d) of Theorem (8.4.4) imply that $|x| > 0$ as well, and since $|B| \leq A$ it follows from (8.1.15) and the fact that $|B|\,|x| = A|x|$ that $|B| = A$. If we define $\theta_k \in \mathbf{R}$ by $e^{i\theta_k} = x_k/|x_k|$, $k = 1, \ldots, n$, if $\lambda \equiv e^{i\varphi}\rho(A)$, and if we set $D \equiv \operatorname{diag}(e^{i\theta_1}, \ldots, e^{i\theta_n})$, then $x = D|x|$ and $\lambda x = e^{i\varphi}\rho(A)D|x| = BD|x| = Bx$. Thus, $e^{-i\varphi}D^{-1}BD|x| = \rho(A)|x| = A|x|$. This identity, together with the fact that $|x| > 0$ and $|e^{-i\varphi}D^{-1}BD| = A$, implies that $e^{-i\varphi}D^{-1}BD = A$. □

Exercise. Supply the details for the last part of the preceding proof. *Hint:* Let $C = e^{-i\varphi}D^{-1}BD$ and observe that

$$A|x| = C|x| = |C|x|| \leq |C|\,|x| = A|x|$$

so that equality holds in the triangle inequality, $\arg(c_{ij}|x_j|) = \text{constant}$, $c_{ij} \geq 0$, and $c_{ij} = a_{ij}$.

When $A > 0$, we know from Perron's theorem that $\rho(A)$ is the unique eigenvalue of A of largest modulus. When $A \geq 0$ there may be more than one eigenvalue of maximum modulus, but in this case A must have a special form and these eigenvalues must be located in a very regular pattern.

8.4.6 **Corollary.** Let $A \in M_n$, suppose A is nonnegative and irreducible, and suppose the set $S = \{\lambda_n = \rho(A),\ \lambda_{n-1}, \ldots, \lambda_{n-k+1}\}$ of eigen-

values of maximum modulus $\rho(A)$ has exactly k distinct elements. Then each eigenvalue $\lambda_i \in S$ has algebraic multiplicity 1 and

$$S = \{e^{2\pi i p/k}\rho(A): p = 0, 1, \ldots, k-1\}$$

that is, these maximum modulus eigenvalues are precisely the k kth roots of unity times $\rho(A)$. Moreover, if λ is *any* eigenvalue of A, then $e^{2\pi i p/k}\lambda$ is an eigenvalue for all $p = 0, 1, \ldots, k-1$.

Proof: For each eigenvalue in S, write $\lambda_{n-p} = e^{i\varphi_p}\rho(A)$, $p = 0, 1, \ldots, k-1$, that is, $\varphi_p \equiv \arg(\lambda_{n-p})$. Assume that $k > 1$, relabel the eigenvalues if necessary, and redefine the arguments if necessary so that $0 = \varphi_0 < \varphi_1 < \varphi_2 < \cdots < \varphi_{k-1} < 2\pi$. Applying the preceding theorem with $B \equiv A$ and $\lambda \equiv \lambda_{n-p}$, we find that $A = B = e^{i\varphi_p}D_p A D_p^{-1}$ for $p = 0, 1, \ldots, k-1$. Since $D_p A D_p^{-1}$ is similar to A, it has the same eigenvalues as A, so this identity shows that the set of eigenvalues of A is carried into itself if it is rotated in the complex plane by the angle φ_p for any $p = 0, 1, \ldots, k-1$; this is the last assertion above (provided we show that $\varphi_p = 2\pi p/k$). Furthermore, $\lambda_n = \rho(A)$ is known to be an algebraically simple eigenvalue of A (since A is irreducible), so letting $p = 1, 2, \ldots, k-1$ in succession, we conclude that all λ_{n-p} are algebraically simple as well.

Slightly more can be said, however. Since $S = \{\lambda_n, \lambda_{n-1}, \ldots, \lambda_{n-k+1}\} = \{e^{i\varphi_p}\lambda_n, e^{i\varphi_p}\lambda_{n-1}, \ldots, e^{i\varphi_p}\lambda_{n-k+1}\}$ for each $p = 0, 1, \ldots, k-1$, there must be some $q = q(p)$ such that $\rho(A) = \lambda_n = e^{i\varphi_p}\lambda_q$; that is, for each p there is some $q = q(p)$ such that $\varphi_p = 2\pi - \varphi_q$ [i.e., $\varphi_p = -\varphi_q \pmod{2\pi}$], and so $e^{-i\varphi_p}\rho(A) \in S$. Moreover, if we iterate the representation for A given by Theorem (8.4.5) and take $B \equiv e^{i\varphi_r}D_r A D_r^{-1}$ and $\lambda \equiv \lambda_{n-m} = e^{i\varphi_m}\rho(A)$ for any choice of r, m with $0 \le r, m \le k-1$, we find that

$$A = e^{i\varphi_r}D_r\{e^{i\varphi_m}D_m A D_m^{-1}\}D_r^{-1} = e^{i(\varphi_m+\varphi_r)}D_r D_m A (D_r D_m)^{-1}$$

so that, by the same argument as above, the set of eigenvalues of A is rotated into itself by a rotation of angle $\varphi_m + \varphi_r$ in the complex plane. In particular, $\lambda_n e^{i(\varphi_m+\varphi_r)} = e^{i(\varphi_m+\varphi_r)}\rho(A)$ must be an eigenvalue of A (of maximum modulus), so for some $j = j(m, r)$ we must have $\varphi_m + \varphi_r = \varphi_j \pmod{2\pi}$.

Consider the set $G \equiv \{\varphi_0 = 0, \varphi_1, \ldots, \varphi_{n-k+1}\} \subset [0, 2\pi)$. The preceding paragraph contains the information that (a) $0 \in G$; (b) if $\varphi_i, \varphi_j \in G$, then $\varphi_i + \varphi_j \pmod{2\pi} \in G$; (c) if $\varphi_i \in G$, then $-\varphi_i \pmod{2\pi} \in G$. Moreover, it is clear that (d) if $\varphi_i, \varphi_j \in G$, then $\varphi_i + \varphi_j = \varphi_j + \varphi_i \pmod{2\pi}$. Thus, G is an Abelian group with exactly k elements; the group operation is "addition modulo 2π." Since the order of any element of a finite Abelian group must divide the order of the group, every $e^{i\varphi_m}$ must be a pth root of unity, where $p = p(m)$ is a divisor of k. We can prove this, and more, with a direct argument that makes no use of group theory.

Since $\varphi_m + \varphi_r = \varphi_j \pmod{2\pi}$ for some j for each r and m, by induction we find that (setting $r = m = 1$) $r\varphi_1 \in G$ for all $r = 0, 1, 2, \ldots \pmod{2\pi}$; that is, $e^{ir\varphi_1}\rho(A) \in S$ for all $r = 0, 1, 2, \ldots$. But if $e^{i\varphi_1}$ were *not* a root of unity, this would imply that there are infinitely many distinct elements of S, which is absurd. Thus, $e^{ip\varphi} = 1$ for some p with $1 < p \le k$; we may assume that p is the smallest such integer for which this is correct. Recall that $0 < \varphi_1 < \varphi_2 < \cdots < \varphi_{n-k+1} < 2\pi$, and for some fixed index m consider φ_m. The interval $[0, 2\pi)$ is divided into exactly p half-open subintervals (open at the right) of length $2\pi/p$ by the $p+1$ points $0, \varphi_1, 2\varphi_1, \ldots, (p-1)\varphi_1, p\varphi_1 = 2\pi$; the point φ_m must lie in one of these subintervals. Thus, there is some q with $0 \le q \le p-1$ such that $q\varphi_1 \le \varphi_m < (q+1)\varphi_1$; that is, $0 \le \varphi_m - q\varphi_1 < \varphi_1$. Therefore, we must have $\varphi_m - q\varphi_1 = \varphi_j$ for some $j = j(m)$ because we have already shown that if $e^{i\varphi_1}\rho(A)$ is an eigenvalue, then so are $e^{-i\varphi_1}\rho(A)$, $e^{-iq\varphi_1}\rho(A)$, and $e^{-iq\varphi_1 + i\varphi_m}\rho(A)$. But then $0 \le \varphi_m - q\varphi_1 = \varphi_j < \varphi_1$ and φ_1 was chosen to be the *least* nonzero argument, so we must have $\varphi_m - q\varphi_1 = 0$. This shows that *every* argument φ_j is *some* multiple of φ_1, so it must be that $p = k$; that is, $\varphi_1 = 2\pi/k$, since if $p < k$, there would be fewer than k distinct elements in the set $\{e^{i\varphi_1}\rho(A), e^{2i\varphi_1}\rho(A), e^{3i\varphi_1}\rho(A), \ldots\}$ which, nevertheless, must equal all of S. Finally, since each φ_m is some multiple of $\varphi_1 = 2\pi/k$ and since there are k distinct φ_i terms and k distinct multiples of φ_1, it must be that $\varphi_m = m\varphi_1$ for all $m = 0, 1, 2, \ldots, k-1$.

The entire argument has proceeded on the assumption that $k > 1$, but if $k = 1$ the assertions made are trivial. □

8.4.7 **Remark.** If $A \ge 0$ is irreducible and has $k > 1$ eigenvalues of maximum modulus, then each nonzero eigenvalue of A lies on a circle centered at 0 in **C** that passes through exactly k eigenvalues of A, all equally spaced around the circle. In particular, k must be a divisor of the number of nonzero eigenvalues of A. Thus, if A is a nonsingular n-by-n nonnegative irreducible matrix with n a prime, there must be either one or n eigenvalues of maximum modulus; there are no other possibilities.

8.4.8 **Corollary.** Suppose $A \in M_n$ is nonnegative and irreducible, and denote $A^m \equiv [a_{ij}^{(m)}]$ for $m = 1, 2, \ldots$. If there are precisely $k > 1$ eigenvalues of A of maximum modulus, then $a_{ii}^{(m)} \equiv 0$ for all $i = 1, 2, \ldots, n$ whenever m is not an integral multiple of k. In particular, all $a_{ii} = 0$.

Proof: Use Corollary (8.4.6) to choose an eigenvalue $\lambda = e^{i\varphi}\rho(A)$ of A of maximum modulus with $\varphi = 2\pi/k$. Thus, $e^{im\varphi}$ is not real and positive whenever m is not an integral multiple of k. Using Theorem (8.4.5) with $B = A$ and $\lambda = e^{i\varphi}\rho(A)$, we find that $A = e^{i\varphi}DAD^{-1}$, so $A^m = e^{im\varphi}DA^mD^{-1}$ and $a_{ii}^{(m)} = e^{im\varphi}a_{ii}^{(m)}$ for all $i = 1, \ldots, n$ and all $m = 1, 2, 3, \ldots$. If $e^{im\varphi}$ is not real and positive, this is impossible if $a_{ii}^{(m)} > 0$, so we must have $a_{ii}^{(m)} \equiv 0$ for all $i = 1, \ldots, n$ whenever m is not a multiple of k. □

Exercise. Suppose that $A \in M_n$ is nonnegative and irreducible. Show that in order to guarantee that $\rho(A)$ is the unique eigenvalue of A of maximum modulus, it is *sufficient* to have some $a_{ii} \neq 0$. However, consider the matrix

$$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

and show that this condition is not *necessary*. Can you find a 2-by-2 counterexample?

8.4.9 **Remark.** A result sharper than Corollary (8.4.8) is true: If $A \geq 0$ is irreducible and if A has $k > 1$ eigenvalues of maximum modulus, then there is a permutation matrix P such that

$$PAP^T = \begin{bmatrix} 0 & A_{12} & & \mathbf{0} \\ \vdots & 0 & \ddots & \\ 0 & & \ddots & A_{k-1,k} \\ A_{k,1} & 0 & \cdots & 0 \end{bmatrix}$$

where the k main diagonal zero blocks are square and the blocks A_{ij} shown are not necessarily zero. In particular, all the main diagonal entries a_{ii} must vanish [Var, p. 28].

Although the hypothesis of irreducibility is essential to get the regular pattern of maximum modulus eigenvalues described in Corollary (8.4.6), one can get some information in the general case.

8.4.10 **Corollary.** If $A \in M_n$ and $A \geq 0$, if $\rho(A) > 0$, and if λ is an eigenvalue of A such that $|\lambda| = \rho(A)$, then $\lambda/\rho(A) = e^{i\varphi}$ is a root of unity, $e^{ik\varphi} = 1$ for some k with $1 \leq k \leq n$, and $e^{ip\varphi}\rho(A)$ is an eigenvalue of A for all $p = 0, 1, 2, \ldots, k-1$.

Proof: If A is irreducible, then the assertion follows from Corollary (8.4.6). If A is not irreducible, then by a permutation similarity, A can be brought into the block upper triangular form

$$\begin{bmatrix} A_1 & & & * \\ & A_2 & & \\ & & \ddots & \\ \mathbf{0} & & & A_r \end{bmatrix}$$

where each A_j is a square matrix that is either irreducible or zero. The eigenvalues of A are the union of the eigenvalues of the diagonal block matrices $A_1, \ldots, A_r$, and the structure of the set of maximum modulus eigenvalues of each A_j is given by Corollary (8.4.8). □

Exercise. Let

$$A_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad A_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

and consider

$$A = \begin{bmatrix} A_1 & * \\ 0 & A_2 \end{bmatrix} \in M_5$$

to show that for general nonnegative A there can be more eigenvalues of maximum modulus than just $\rho(A)$ and a single set of rotations of $\rho(A)$ through powers of a single root of unity.

Problems

1. Show by examples that the items in Perron's theorem (8.2.11) that are not included in Theorem (8.4.4) are not generally true of irreducible nonnegative matrices.

2. Show by example that $\rho(I+A) = 1 + \rho(A)$ is not true for all $A \in M_n$. Give a necessary and sufficient condition on A for this identity to be correct.

3. Irreducibility is a sufficient but not necessary condition that a nonnegative matrix have a positive eigenvector. Consider $\left[\begin{smallmatrix} 1 & 1 \\ 0 & 0 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 1 & 0 \\ 1 & 0 \end{smallmatrix}\right]$ to show that a reducible nonnegative matrix may or may not have a positive eigenvector.

4. If $A \geq 0$ is irreducible, show that the entries of the matrices $[\rho(A)^{-1}A]^m$ are uniformly bounded as $m \to \infty$.

5. We have shown that an irreducible matrix has a positive Perron vector. Suppose that $A \geq 0$, $\rho(A) > 0$, $x \geq 0$, $x \neq 0$, and $Ax = \rho(A)x$; show that if x is not positive, then A is reducible. If x is positive, must A be irreducible?

6. Suppose $A \geq 0$ is irreducible and that $B \geq 0$ commutes with A. If x is the Perron vector of A, show that $Bx = \rho(B)x$. *Hint:* Theorem (8.4.4d).

7. Show that the companion matrix of the polynomial $x^k - 1 = 0$ is an example of a k-by-k nonnegative matrix with k eigenvalues of maximum modulus. Sketch the location of these eigenvalues in the complex plane.

8. Let $k_1, k_2, \ldots, k_p$ be given positive integers. Show how to construct a single nonnegative matrix whose eigenvalues of maximum modulus are precisely the k_1 (k_1)th roots of unity, the k_2 (k_2)th roots of unity, ..., and the k_p (k_p)th roots of unity.

9. Explain why an irreducible nonnegative matrix A is said to be *cyclic of index* k if it has $k \geq 1$ eigenvalues of maximum modulus.

10. If $A \geq 0$ is irreducible and is cyclic of index $k \geq 1$, show that the characteristic polynomial $p_A(t) = t^r(t^k - \rho(A)^k)(t^k - \mu_2^k)\cdots(t^k - \mu_m^k)$ for some $r, m \geq 0$ and some complex numbers μ_i with $|\mu_i| < \rho(A)$, $i = 2, \ldots, m$. Comment on the pattern of zero and nonzero coefficients in $p_A(t)$ and give a criterion for A to have only one eigenvalue of maximum modulus based on the form of the characteristic polynomial. *Hint:* In the proof of Corollary (8.4.6) we found that if $\varphi = 2\pi/k$ and if λ is an eigenvalue of A, then so are $e^{ir\varphi}\lambda$, $r = 0, 1, 2, \ldots$.

11. Let $n > 1$ be a prime number. Show that if $A \in M_n$ is nonnegative, irreducible, and nonsingular, either $\rho(A)$ is the only eigenvalue of A of maximum modulus or all the eigenvalues of A have maximum modulus.

12. Consider $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ and show that the conclusion of Corollary (8.4.8) cannot be improved in general to assert that the main diagonals of all powers of A must vanish.

13. If $A \geq 0$, show that irreducibility of A depends only the *location* of the zero entries and not on the magnitude of the nonzero entries.

14. If $A, B \in M_n$, then AB and BA have the same set of eigenvalues. Consider $\begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$ and $\begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}$ to show that even if A and B are nonnegative, it is possible to have AB irreducible and BA reducible. This example shows that an irreducible matrix can be similar (even unitarily equivalent) to a reducible matrix; explain why. It also shows that no condition involving only the eigenvalues of a matrix can be a definitive test for irreducibility.

15. Let $A \in M_n$ be a given irreducible nonnegative matrix. Show that $A + B$ is irreducible whenever $B \in M_n$ is nonnegative, and that $\rho(A+B) > \rho(A)$ whenever $B \geq 0$ and $B \neq 0$. This is an improvement of (8.1.18) to strict monotonicity, but with the additional assumption of irreducibility. *Hint:* (8.1.18) says that $\rho(A+B) \geq \rho(A)$. If equality holds, use (8.4.5) to show that $B = 0$.

16. Show that (8.4.1) can be sharpened in the following way. Let $A \in M_n$ be nonnegative and let the minimal polynomial of A have degree m. Show that A is irreducible if and only if $(I+A)^{m-1} > 0$. *Hint:* Consider $I + A + A^2 + \cdots + A^{m-1} + A^m + \cdots + A^{n-1}$ and use the minimal polynomial to express A^m and higher powers in terms of $I, A, \ldots, A^{m-1}$.

17. Let $A \in M_n$ be a given nonnegative matrix and consider the problem of finding a best rank 1 approximation to A in the sense of least squares; that is, find a rank 1 $X \in M_n$ such that $\|A - X\|_2 = \min\{\|A - Y\|_2 : Y \in M_n$ has rank $1\}$. Suppose that the Perron root of AA^T is simple, which would be the case if either AA^T or A^TA is irreducible. Why? Show that such a best X is nonnegative, unique, and given by $X = \sqrt{r}vw^T$, where $r = \rho(AA^T)$ is the Perron root of AA^T, and $v, w \in \mathbf{R}^n$ are nonnegative unit vectors that are, respectively, unit eigenvectors of AA^T and A^TA associated with the eigenvalue r. *Hint:* Use the characterization of a best rank 1 approximation given in (7.4.1). Notice that AA^T and A^TA are both real symmetric positive semidefinite matrices, so the computation of r, v, and w is, in principle, not too difficult.

18. Use Problem 17 to find a best rank 1 least-squares approximation to each of the matrices

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}$$

Show that a best rank 1 least-squares approximation to $A = I \in M_n$ is not unique; $X = vv^*$ is a best rank 1 least-squares approximation to I for any unit vector $v \in \mathbf{C}^n$.

8.5 Primitive matrices

In practice, the result in Perron's theorem that may have the most frequent application is the limit statement in Theorem (8.2.8). An examination of Theorem (8.4.4) shows that the only hypothesis lacking for an

application of Lemma (8.2.7) to irreducible matrices is the condition that the spectral radius is the only eigenvalue of maximum modulus. Since $A=\begin{bmatrix}0&1\\1&0\end{bmatrix}$ is an example of a nonnegative irreducible matrix with two eigenvalues of maximum modulus (and for which $\lim_{m\to\infty}A^m$ does not exist), some further restriction of the class of irreducible matrices is necessary; the most economical procedure is to assume exactly what we need.

8.5.0 **Definition.** A nonnegative matrix $A\in M_n$ is said to be *primitive* if it is irreducible and has only one eigenvalue of maximum modulus.

The notion of primitivity is due to Frobenius (1912). The limit result now follows directly from Lemma (8.2.7) with the same proof as for Theorem (8.2.8).

8.5.1 **Theorem.** If $A\in M_n$ is nonnegative and primitive, then

$$\lim_{m\to\infty}[\rho(A)^{-1}A]^m=L>0$$

where $L=xy^T$, $Ax=\rho(A)x$, $A^Ty=\rho(A)y$, $x>0$, $y>0$, and $x^Ty=1$. Moreover, if λ_{n-1} is an eigenvalue of A such that $|\lambda|\le|\lambda_{n-1}|$ for every eigenvalue $\lambda\ne\rho(A)$, and if $|\lambda_{n-1}|/\rho(A)<r<1$, then there exists a constant $C=C(r,A)$ such that $\|[\rho(A)^{-1}A]^m-L\|_\infty\le Cr^m$ for all $m=1,2,\ldots$. □

We have now generalized all of Perron's theorem from the class of positive matrices to the class of primitive nonnegative matrices. In practice, however, one still faces the problem of testing a given nonnegative matrix for primitivity; ideally, one would hope to be able to do so without explicit calculation of the eigenvalues. The following characterization of primitivity, while not itself a computationally effective test, leads to several useful criteria.

8.5.2 **Theorem.** If $A\in M_n$ is nonnegative, then A is primitive if and only if $A^m>0$ for some $m\ge1$.

Proof: If $A\ge0$ and $A^m>0$, then from every node P_i of the directed graph $\Gamma(A)$ of A to every other node P_j there must be a directed path of exact length m (Corollary 6.2.18). Since this is a stronger property than irreducibility, A must be irreducible. Application of Perron's theorem (8.2.11d, e) to A^m, as in (8.4.3), then implies that A must be primitive. Conversely, if A is primitive, then $\lim_{m\to\infty}[\rho(A)^{-1}A]^m=L>0$ by Theorem (8.5.1), and so for some $m\ge1$ it must be that $[\rho(A)^{-1}A]^m>0$. □

This characterization together with the very sharp information we have about the maximum modulus eigenvalues of nonnegative irreducible matrices now gives us a graphical criterion for primitivity reminiscent of the graphical criterion for irreducibility. Recall that the greatest common divisor (g.c.d.) of a sequence of positive integers $k_1,k_2,\ldots$ is the largest integer $k\ge1$ such that k is a divisor of all $k_1,k_2,\ldots$.

8.5.3 **Theorem.** Let $A\in M_n$ be nonnegative and irreducible, and let $\{P_i\}$ denote the set of nodes of the directed graph $\Gamma(A)$. Denote by $L_i=\{k_1^{(i)},k_2^{(i)},\ldots\}$ the set of lengths of all directed paths in $\Gamma(A)$ that both start and end at the node P_i, $i=1,2,\ldots,n$. Denote by g_i the greatest common divisor of all the lengths in L_i. Then A is primitive if and only if all $g_i=1$, $i=1,2,\ldots,n$.

Proof: Observe that no set of lengths L_i is empty since A is irreducible; for each i and for any $j\ne i$ there is a path in $\Gamma(A)$ joining P_i to P_j and there is also a path in $\Gamma(A)$ joining P_j to P_i. If A is primitive, then by Theorem (8.5.2) there is some $m\ge1$ such that $A^m>0$, and hence $A^k>0$ for all $k\ge m$. But then $m,m+1,m+2,\ldots\in L_i$ for all $i=1,\ldots,n$, and hence $g_i=1$ for all $i=1,\ldots,n$.

Now suppose $A=[a_{ij}]$ is not primitive. If A has exactly $k>1$ eigenvalues of maximum modulus, then by Corollary (8.4.8) we know that $a_{ii}^{(m)}\equiv0$ for all $i=1,\ldots,n$ and for all m such that m is not an integral multiple of k. Thus, $L_i\subset\{k,2k,3k,\ldots\}$, and hence $g_i\ge k>1$ for all $i=1,\ldots,n$. □

8.5.4 **Remark.** Somewhat more than the assertions in Theorem (8.5.3) is true; in fact, $g_1=g_2=\cdots=g_n$ always, and the common value of the g_i terms is precisely the number of eigenvalues of A of maximum modulus. This is a theorem of Romanovsky.

The following result is useful in many situations; in particular, it shows that an irreducible nonnegative matrix with positive main diagonal must be primitive.

8.5.5 **Lemma.** If $A\in M_n$ is nonnegative and irreducible, and if all the main diagonal entries of A are positive, then $A^{n-1}>0$.

Proof: If $\alpha=\min\{a_{11},a_{22},\ldots,a_{nn}\}$, and if we define

$$B=A-\operatorname{diag}(a_{11},a_{22},\ldots,a_{nn})$$

then B is nonnegative and irreducible (because A is irreducible), and $A \ge \alpha I + B = \alpha[I + (1/\alpha)B]$ and hence $A^{n-1} \ge \alpha^{n-1}[I + (1/\alpha)B]^{n-1} > 0$ by Lemma (8.4.1). □

Exercise. Note that as a nonnegative square matrix with positive diagonal entries is powered, any entry that becomes positive remains positive in all successive powers.

Although an irreducible matrix may have a reducible power, all powers of a primitive matrix are primitive.

8.5.6 **Lemma.** Let $A \in M_n$ be nonnegative and primitive. Then A^k is nonnegative, irreducible, and primitive for all $k = 1, 2, \ldots$.

Proof: Since all sufficiently large powers of A are positive, the same is true for A^k for any k. If A^k were reducible for some k, then all powers of A^k would also be reducible and hence could not be positive. Since this contradicts the fact that all sufficiently large powers of A are positive, it is impossible for any power of A to be reducible. □

The characterization in Theorem (8.5.2) is not in itself a computationally effective test for primitivity since no upper bound on the powers to be computed is given. If one finds an m such that $A^m > 0$, then A is primitive; but when does one stop computing if one has not yet found a positive power? A finite bound which answers this question is given by the following theorem.

8.5.7 **Theorem.** Let $A \in M_n$ be nonnegative. If A is primitive, then $A^k > 0$ for some positive integer $k \le (n-1)n^n$.

Proof: Because A is irreducible, there is a directed path from the node P_1 in $\Gamma(A)$ back to node P_1; the shortest such path has length $k_1 \le n$. The matrix A^{k_1} therefore has a positive entry in its 1,1 position, and any power of A^{k_1} will also have a positive 1,1 entry. Because A is primitive, A^{k_1} must be irreducible by Lemma (8.5.6), and so there is a directed path from the node P_2 in $\Gamma(A^{k_1})$ back to the node P_2; the shortest such path has length $k_2 \le n$. The matrix $(A^{k_1})^{k_2} = A^{k_1 k_2}$ therefore has positive 1,1 and 2,2 entries. This process can be continued down the main diagonal until we obtain a matrix $A^{k_1 k_2 \cdots k_n}$ (with each $k_i \le n$), which is irreducible and has positive diagonal entries, and hence $[A^{k_1 k_2 \cdots k_n}]^{n-1} > 0$ by Lemma (8.5.5). Since

$$k_1 k_2 \cdots k_n (n-1) \le n \cdot n \cdots n(n-1) = n^n(n-1)$$

we are done. □

If A is a given primitive matrix, the least k such that $A^k > 0$ is called the *index of primitivity* of A and is usually denoted by $\gamma(A)$. We have seen that $\gamma(A) \le n-1$ if A has a positive diagonal and $\gamma(A) \le n^n(n-1)$ in general. The latter bound can be improved considerably.

8.5.8 **Theorem.** Let $A \in M_n$ be a nonnegative primitive matrix, and suppose the shortest simple directed cycle in $\Gamma(A)$ has length s. Then $A^{n+s(n-2)} > 0$, that is, $\gamma(A) \le n + s(n-2)$.

Proof: Because A is irreducible, every node in $\Gamma(A)$ lies on a cycle and the shortest cycle from any node back to itself will be a simple cycle of length at most n. By a permutation, we may assume that the nodes in the shortest such cycle are $P_1, P_2, \ldots, P_s$. Notice that $n + s(n-2) = n - s + s(n-1)$ and consider $A^{n-s+s(n-1)} = A^{n-s}(A^s)^{n-1}$. Write A^{n-s} in block form

$$A^{n-s} = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}$$

with $X_{11} \in M_s$ and $X_{22} \in M_{n-s}$. Then X_{11} has at least one nonzero entry in each row because the nodes $P_1, \ldots, P_s$ comprise a cycle in $\Gamma(A)$ and hence from each node P_i in the graph $\Gamma(A_{n-s})$ there is some arc to some node P_j (perhaps $P_i = P_j$) in $\Gamma(A^{n-s})$; this is correct if $1 \le i, j \le s$. There is at least one nonzero entry in each row of X_{21} because for each node $P_{s+1}, \ldots, P_n$ not in the cycle there must be a directed path in $\Gamma(A)$ of length not more than $n-s$ (the number of nodes not in the cycle) to some node in the cycle. By then going a sufficient number of additional steps around the cycle, it is clear that there is a directed path of length exactly $n-s$ in $\Gamma(A)$ from every node not in the cycle to some node in the cycle.

Now write $(A^s)^{n-1}$ in block form as

$$(A^s)^{n-1} = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}$$

where $Y_{11} \in M_s$ and $Y_{22} \in M_{n-s}$. Because $P_1, \ldots, P_s$ comprise a cycle in $\Gamma(A)$, there is a loop at each node $P_1, \ldots, P_s$ in $\Gamma(A^s)$. Since A is primitive, A^s is also primitive, and hence is irreducible. From each node $P_1, \ldots, P_s$ in $\Gamma(A^s)$ there is a path in $\Gamma(A^s)$ of length at most $n-1$ to any other node. By first going a sufficient number of times around the loop at the starting node, we can always construct such a path of length exactly $n-1$. This shows that $Y_{11} > 0$ and $Y_{12} > 0$.

To complete the argument we compute

$$A^{n-s}(A^s)^{n-1} = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} \geq \begin{bmatrix} X_{11}Y_{11} & X_{11}Y_{12} \\ X_{21}Y_{11} & X_{21}Y_{12} \end{bmatrix}$$

Because each row of the X blocks in the last expression contains at least one nonzero entry, and because each of the Y blocks in the last expression is positive, the entire block matrix is positive and $A^{n-s}(A^s)^{n-1} > 0$. □

One consequence of (8.5.8) is a celebrated result of H. Wielandt, which gives a sharp upper bound for the index of primitivity of a general primitive matrix.

8.5.9 **Corollary.** If $A \in M_n$ is a nonnegative matrix, then A is primitive if and only if $A^{n^2-2n+2} > 0$.

Proof: If any power of A is positive, then A is primitive, so only the converse implication is of interest. If $n=1$, the result is trivial, so assume $n>1$. If A is primitive, then it is irreducible and there are cycles in $\Gamma(A)$. If the shortest cycle in $\Gamma(A)$ had length n, then the length of every cycle in $\Gamma(A)$ is a multiple of n and hence A could not be primitive by Theorem (8.5.3). Thus, the length of the shortest cycle in $\Gamma(A)$ is $n-1$ or less, and hence by Theorem (8.5.8) we have

$$\gamma(A) \leq n+s(n-2) \leq n+(n-1)(n-2) = n^2-2n+2 \qquad □$$

Wielandt gave an example (see Problem 4 at the end of this section) to show that the bound $\gamma(A) \leq n^2-2n+2$ is the best possible bound for matrices that have all diagonal entries 0. We know that if all main diagonal entries are positive, then A is primitive if and only if $A^{n-1} > 0$. The following result of Holladay and Varga uses the same ideas employed in the proof of Theorem (8.5.8) to provide a bound on the index of primitivity if some, but perhaps not all, of the main diagonal entries are positive.

8.5.10 **Theorem.** Let $A \in M_n$ be nonnegative and irreducible, and suppose A has d positive main diagonal entries, $1 \leq d \leq n$. Then $A^{2n-d-1} > 0$; that is, $\gamma(A) \leq 2n-d-1$.

Proof: Under the hypotheses stated, A must be primitive, and the minimum length cycle in $\Gamma(A)$ has length 1. In fact, there are d such cycles. By a permutation, we may assume that $P_1, \ldots, P_d$ are the nodes in $\Gamma(A)$ that have loops. Consider $A^{2n-d-1} = A^{n-d}(A^1)^{n-1}$ and write

$$A^{n-d} = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}, \qquad A^{n-1} = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}$$

where $X_{11}, Y_{11} \in M_d$ and $X_{22}, Y_{22} \in M_{n-d}$. The same arguments used in the proof of Theorem (8.5.8) to treat the correspondingly placed blocks of A^{n-s} and $(A^s)^{n-1}$ show that each row of the blocks X_{11} and X_{21} contains at least one nonzero entry, and the blocks Y_{11} and Y_{12} are positive. It follows that the product $A^{n-d}A^{n-1}$ is positive by the same reasoning used in the Theorem (8.5.8). □

Exercise. Show that the matrix $A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$ is primitive. What are its eigenvalues? Compute the bounds on $\gamma(A)$ given by (8.5.9) and (8.5.10). What is the exact value of $\gamma(A)$?

As a final remark, we note that if one wishes to verify that a given nonnegative matrix is primitive, then one could check that the matrix is irreducible and that Wielandt's condition (8.5.9) is met. Matrices arising in practice frequently have a special structure that makes it easy to see whether or not the associated directed graph is strongly connected. Furthermore, if the matrix is irreducible and any main diagonal entry is positive, then it must be primitive. However, if the matrix is large and there is no special structure or symmetry to its entries, or if all the main diagonal entries are zero, then it may be necessary to use Lemma (8.4.1) or Corollary (8.5.9) to check irreducibility or primitivity. In either case, the required number of matrix multiplications will be considerably reduced if the matrix in question is squared repeatedly until the resulting power exceeds the critical value ($n-1$ or n^2-2n+2, respectively). For example, if $n=10$, then calculation of $(I+A)^2$, $(I+A)^4$, $(I+A)^8$, and $(I+A)^{16}$ is sufficient to verify irreducibility; this is 4 matrix multiplications instead of the 8 required by a direct application of Lemma (8.4.1). Similarly, if A is nonnegative, then calculation of A^2, A^4, A^8, A^{16}, A^{32}, A^{64}, and A^{128} is sufficient to verify primitivity; this is 7 matrix multiplications instead of 81. Note that we are making implicit use of Problem 3 in these considerations.

Problems

1. Write out the proof of Theorem (8.5.1).

2. If $A \in M_n$ is nonnegative and primitive, show that $\lim_{m\to\infty}[a_{ij}^{(m)}]^{1/m} = \rho(A)$ for all $i, j = 1, \ldots, n$. Compare this result with Corollary (5.6.14). Can either part of the hypothesis of primitivity be omitted?

3. Show that if $A \geq 0$ and $A^k > 0$, then $A^m > 0$ for all $m \geq k$. If A is primitive, show that A^k is primitive for any positive integer k. However, if A and B are both primitive, it could be that AB is not primitive. *Hint:* Consider $\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$.

4. Use $\Gamma(A)$ to show that Wielandt's matrix

$$A = \begin{bmatrix} 0 & 1 & & & 0 \\ 0 & & 1 & & \\ \vdots & & \ddots & \ddots & \\ 0 & 0 & & & 1 \\ 1 & 1 & 0 & \cdots & 0 \end{bmatrix} \in M_n$$

is irreducible and primitive for all $n \geq 3$. Then show that the $(1,1)$ entry of A^{n^2-2n+1} is zero but $A^{n^2-2n+2} > 0$. *Hint:* Think of A as a linear transformation acting on the standard basis $\{e_1, \ldots, e_n\}$. Then $A: e_i \to ?$ $A^{n-1}: e_i \to ?$ $A^{(n-1)(n-1)}: e_1 \to ?$

5. Let $A \in M_n$ be nonnegative and irreducible. Show that A is primitive if at least one main diagonal entry is positive. Show that this sufficient condition is necessary for $n=2$ but not for $n \geq 3$.

6. Let $A = [a_{ij}] \in M_n$ be nonnegative, and suppose $a_{kk} > 0$ for some $k = 1, 2, \ldots, n$. Show that the k,k entry of every power of A is also positive. If $a_{kk} = 0$ but the k,k entry of A^2 is positive, is the k,k entry of A^3 positive?

7. Justify in detail the computational shortcuts suggested at the end of this section.

8. If A is any idempotent matrix, then $A = \lim_{m\to\infty} A^m$. Show that if A is nonnegative, irreducible, and idempotent, then A is a positive matrix of rank 1.

9. Give an example to show that $\lim_{m\to\infty}[\rho(A)^{-1}A]^m$ can exist even if $A \geq 0$ is not primitive. Indeed, A can be reducible and can also have multiple eigenvalues of maximum modulus.

10. Prove the following partial converse of Theorem (8.5.1): If $A \in M_n$ is nonnegative and irreducible, and if $\lim_{m\to\infty}[\rho(A)^{-1}A]^m$ exists, then A is primitive. *Hint:* If $|\mu| = \rho(A)$, $\mu \neq \rho(A)$, and $Az = \mu z$, $z \neq 0$, then $[\rho(A)^{-1}A]^m z \to ?$

11. Show that $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ is irreducible, but A^2 is reducible. Does this contradict (8.5.6)?

12. Give an example of an irreducible nonnegative matrix $A \in M_n$ such that $\lim_{m\to\infty}[\rho(A)^{-1}A]^m$ does not exist.

13. If $\epsilon > 0$ and if $A \in M_n$ is nonnegative and irreducible, prove that $A + \epsilon I$ is primitive.

14. A nonnegative matrix $A = [a_{ij}]$ is said to be combinatorially symmetric provided that $a_{ij} > 0$ if and only if $a_{ji} > 0$ for all $i, j = 1, \ldots, n$. Show that if A is combinatorially symmetric and primitive, then $A^{2n-2} > 0$. *Hint:* Consider A^2 and use (8.5.6) and (8.5.10). Can you strengthen the bound for $\gamma(A)$, given more information about the cycle structure of $\Gamma(A)$? *Hint:* Use (8.5.8).

15. Show that if $A \in M_n$ is nonnegative, irreducible, and nonsingular with n a prime number, then either (a) A is primitive or (b) all the eigenvalues of A have maximum modulus and A is similar to the companion matrix of $x^n - \rho(A)^n = 0$.

16. One way to compute the Perron vector and spectral radius of a nonnegative matrix $A \in M_n$ is the power method:

$$x^{(0)} \text{ is an arbitrary positive vector,} \qquad \sum_{i=1}^{n} x_i^{(0)} = 1$$

$$y^{(m+1)} = Ax^{(m)} \quad \text{for all} \quad m = 0, 1, 2, \ldots$$

$$x^{(m+1)} = \frac{y^{(m+1)}}{\sum_{i=1}^{n} y_i^{(m+1)}} \quad \text{for all} \quad m = 0, 1, 2, \ldots$$

If A is primitive, show that the sequence of vectors $x^{(m)}$ converges to the (right) Perron vector of A and that the sequence of numbers $\sum_{i=1}^{n} y_i^{(m+1)}$ converges to the Perron root of A. What is the rate of convergence? Is the hypothesis of primitivity necessary?

17. If $A \in M_n$ is nonnegative, show that primitivity of A depends only on the *location* of the zero entries and not on the magnitude of the nonzero entries.

18. If $A \in M_n$ is nonnegative, irreducible, and symmetric, show that A is primitive if and only if $A + \rho(A)I$ is nonsingular. In particular, this condition is met if A is positive semidefinite. Symmetric nonnegative matrices with 0's and 1's as entries arise naturally as adjacency matrices of undirected graphs.

19. If $A \in M_n$ is primitive and $k \geq \gamma(A)$, show that $A^k > 0$.

20. Provide the details for the proof of Theorem (8.5.10).

21. Calculate the eigenvalues and eigenvectors of each of the following matrices and categorize them according to the key concepts of the chapter (nonnegative, irreducible, primitive, positive, and so forth): $\begin{bmatrix}1&1\\1&1\end{bmatrix}$, $\begin{bmatrix}0&1\\1&1\end{bmatrix}$, $\begin{bmatrix}1&0\\1&1\end{bmatrix}$, $\begin{bmatrix}1&0\\0&1\end{bmatrix}$, $\begin{bmatrix}1&0\\0&0\end{bmatrix}$, $\begin{bmatrix}0&1\\0&0\end{bmatrix}$, $\begin{bmatrix}0&0\\0&0\end{bmatrix}$. These provide a good illustration of the possibilities that can occur.

22. In the proof of Theorem (8.5.8), show that each column of X_{11} and X_{12} contains at least one nonzero entry. Show that $Y_{21}>0$.

Further Reading. For a proof of Romanovsky's theorem mentioned in (8.5.4), see V. Romanovsky, "Recherches sur les Chaines de Markoff," *Acta Math.* 66 (1936), 147-251.

8.6 A general limit theorem

Even if a nonnegative matrix A is irreducible, the normalized powers of A need have no limit, as the example

$$A=\begin{bmatrix}0&1\\1&0\end{bmatrix}$$

easily shows. Nevertheless, there is a precise sense in which, *on the average,* this limit does exist.

8.6.1 **Theorem.** Let $A\in M_n$ be nonnegative and irreducible, let $Ax=\rho(A)x$, $A^Ty=\rho(A)y$, $x^Ty=1$, and $L=xy^T$. Then

$$\lim_{N\to\infty}\frac{1}{N}\sum_{m=1}^{N}[\rho(A)^{-1}A]^m=L$$

Moreover, there exists a finite positive constant $C=C(A)$ such that

$$\left\|\frac{1}{N}\sum_{m=1}^{N}[\rho(A)^{-1}A]^m-L\right\|_\infty\le\frac{C}{N}$$

for all $N=1,2,\ldots$.

Proof: If we set $\lambda=\rho(A)$ and choose for y and x the left and right Perron vectors of A, respectively, then hypotheses (1)-(5) of Lemma (8.2.7) are satisfied and hence the matrix

$$I-[\rho(A)^{-1}A-L]=\rho(A)^{-1}[\rho(A)I-(A-\rho(A)L)]$$

is invertible. Using (e) of Lemma (8.2.7) and the identity in Problem 1 at the end of this section, we compute

$$\frac{1}{N}\sum_{m=1}^{N}[\rho(A)^{-1}A]^m$$

$$=\frac{1}{N}\sum_{m=1}^{N}([\rho(A)^{-1}A-L]^m+L)=L+\frac{1}{N}\sum_{m=1}^{N}[\rho(A)^{-1}A-L]^m$$

$$=L+\frac{1}{N}\{\rho(A)^{-1}A-L\}\{I-[\rho(A)^{-1}A-L]^N\}\{I-[\rho(A)^{-1}A-L]\}^{-1}$$

$$=L+\frac{1}{N}\{\rho(A)^{-1}A-L\}\{I-[\rho(A)^{-1}A]^N+L\}\{I-[\rho(A)^{-1}A-L]\}^{-1}$$

The only part of the second term in this last expression that depends on N is the factor $1/N$ and the term $[\rho(A)^{-1}A]^N$, but the entries of the latter matrix are uniformly bounded as $N\to\infty$ by Corollary (8.1.33). Thus, the second term is of the order of $1/N$ as $N\to\infty$, and hence it tends uniformly to zero. □

An analysis of the hypotheses required by Lemma (8.2.7) and Corollary (8.1.33) shows that exactly the same argument proves the following more general (but less concisely stated) result.

8.6.2 **Theorem.** Let $A\in M_n$ be nonnegative, and let x and y be nonnegative vectors such that $Ax=\rho(A)x$ and $A^Ty=\rho(A)y$. If

(a) $\rho(A)>0$;
(b) $x^Ty>0$;
(c) the matrix

$$I-[\rho(A)^{-1}A-(x^Ty)^{-1}xy^T]$$

is invertible; and
(d) $[\rho(A)^{-1}A]^m$ is uniformly bounded as $m\to\infty$;

then

$$\lim_{N\to\infty}\frac{1}{N}\sum_{m=1}^{N}[\rho(A)^{-1}A]^m=(x^Ty)^{-1}xy^T$$

Moreover, there exists a finite positive constant $C=C(A)$ such that

$$\left\|\frac{1}{N}\sum_{m=1}^{N}[\rho(A)^{-1}A]^m-(x^Ty)^{-1}xy^T\right\|_\infty\le\frac{C}{N}$$

for all $N=1,2,\ldots$.

Problems

1. If $B \in M_n$ and if $I - B$ is invertible, show that

$$\sum_{m=1}^{N} B^m = B(I - B^N)(I - B)^{-1}$$

Hint: Multiply by $I - B$.

2. Prove Theorem (8.6.2).

3. Compare the rate of convergence in Theorems (8.5.1) and (8.6.1). Give an example to show that the rate of convergence in (8.6.1) cannot be improved.

4. Suppose $A \in M_n$ is nonnegative and irreducible, and write $A^m = [a_{ij}^{(m)}]$ for $m = 1, 2, \ldots$. Use Theorem (8.6.1) to show for each given pair (i, j) that $a_{ij}^{(m)} > 0$ for infinitely many values of m. This result may be thought of as a generalization of Theorem (8.5.2). Give an example to show that there may also be infinitely many values of m for which $a_{ij}^{(m)} = 0$.

5. Under the hypotheses of Theorem (8.6.2), show that $a_{ij}^{(m)} > 0$ for infinitely many values of m provided the pair (i, j) is such that $x_i y_j \neq 0$. Why does this result include Problem 4?

6. Show directly that Theorem (8.5.1) implies Theorem (8.6.1) when A is primitive. *Hint:* What is required here is the proof of the following result from analysis: If a sequence is convergent to a finite limit, then it is Cesaro-summable to the same limit.

7. Consider $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, explicitly compute

$$\lim_{N \to \infty} N^{-1} \sum_{m=1}^{N} [\rho(A)^{-1} A]^m$$

compute the value of this limit given by Theorem (8.6.1), and compare.

8.7 Stochastic and doubly stochastic matrices

A nonnegative matrix $A \in M_n$ with the property that all its row sums are $+1$ is said to be a *(row) stochastic matrix* because each row may be thought of as a discrete probability distribution on a sample space with n points. A *column stochastic matrix* is the transpose of a row stochastic matrix; such matrices arose naturally in the intercity population migration model discussed in Section (8.0). Stochastic matrices also arise in the study of Markov chains and in a variety of modeling problems in such fields as economics and operations research.

The set of stochastic matrices in M_n is a compact convex set with a simple but important property. If we denote by $e \in \mathbf{R}^n$ the vector with all components $+1$, a nonnegative matrix $A \in M_n$ is stochastic if and only if $Ae = e$. Thus, the stochastic matrices in M_n form an easily recognized family of nonnegative matrices with a particular positive eigenvector in common. Nonnegative matrices with a positive eigenvector have many special properties [e.g., (8.1.30), (8.1.31), and (8.1.33)] which therefore are possessed by all stochastic matrices.

A stochastic matrix $A \in M_n$ with the property that A^T is also stochastic is said to be *doubly stochastic*; all row and column sums are $+1$. The set of doubly stochastic matrices is also a compact convex set in M_n, and a nonnegative matrix $A \in M_n$ is evidently doubly stochastic if and only if both $Ae = e$ and $e^T A = e^T$. One type of doubly stochastic matrix has already been encountered in (6.3.5), namely an orthostochastic matrix $A \equiv [|u_{ij}|^2]$, where $U = [u_{ij}] \in M_n$ is unitary. That the row and column sums of A are all $+1$ follows from the fact that the rows and columns of U are all Euclidean unit vectors.

Another example of doubly stochastic matrices is the set (group) of permutation matrices. The permutation matrices are really the fundamental and prototypical doubly stochastic matrices, for Birkhoff's theorem says that any doubly stochastic matrix is a convex combination of finitely many permutation matrices. The proof we present for Birkhoff's theorem relies on the fact (see Appendix B) that every point in a compact convex set S is a convex combination of the extreme points of S. We shall show that the extreme points of the set of doubly stochastic matrices are precisely the permutation matrices.

8.7.1 **Theorem** (Birkhoff). A matrix $A \in M_n$ is a doubly stochastic matrix if and only if for some $N < \infty$ there are permutation matrices $P_1, \ldots, P_N \in M_n$ and positive scalars $\alpha_1, \ldots, \alpha_N \in \mathbf{R}$ such that $\alpha_1 + \cdots + \alpha_N = 1$ and $A = \alpha_1 P_1 + \cdots + \alpha_N P_N$.

Proof: The sufficiency of the condition is clear; we must establish its necessity. Let $A = [a_{ij}] \in M_n$ be a given doubly stochastic matrix. If A is a permutation matrix, then there is precisely one entry $+1$ in each row and column and all other entries are 0. If we could write $A = \alpha_1 B + \alpha_2 C$ with $0 < \alpha_1, \alpha_2 < 1$, $\alpha_1 + \alpha_2 = 1$, and B, C doubly stochastic, then every entry of B and C that corresponds to a 0 entry $a_{ij} = 0$ of A must satisfy $0 = a_{ij} = \alpha_1 b_{ij} + \alpha_2 c_{ij}$, so $b_{ij} = c_{ij} = 0$ since α_1 and α_2 are both nonzero and b_{ij}, c_{ij} are nonnegative. Since B and C are doubly stochastic, their row sums are $+1$ and hence nonzero entries must all be $+1$ and in the same positions as the nonzero entries of A; that is, $A = B = C$. This shows that

every permutation matrix is an extreme point of the set of doubly stochastic matrices.

On the other hand, if A is not a permutation matrix there is at least one row of A, say row i, that contains at least two nonzero entries. In that row, choose any nonzero entry $a_{i_1 i_2}$, which must satisfy $0 < a_{i_1 i_2} < 1$ since there are at least two nonzero entries in row i, and the sum of all the (nonnegative) entries in the row is $+1$. Since $0 < a_{i_1 i_2} < 1$ and the sum of all the (nonnegative) entries in column i_2 is $+1$, there must be some other nonzero entry $a_{i_3 i_2}$, $i_3 \neq i_1$, in the same column as $a_{i_1 i_2}$, and $0 < a_{i_3 i_2} < 1$. By the same reasoning, there is some other nonzero entry $a_{i_3 i_4}$, $i_4 \neq i_2$, in the same row as $a_{i_3 i_2}$, and $0 < a_{i_3 i_4} < 1$. If this process is continued and the successive entries chosen in this way are marked, after finitely many steps there will be a first time that an entry a_{ij} is chosen that has previously been chosen. The sequence of entries from the first up to the second occurrence of the entry a_{ij} (including the first but not the second occurrence of a_{ij}) is a finite ordered sequence of entries of A, each successive pair of which is alternately in the same row or column; let $a_{i'j'}$ be the smallest (positive) entry in this sequence. Let $B \in M_n$ be a matrix in which $+1$ occurs in the same position as the first entry a_{ij} of the sequence, -1 occurs in the same position as the second entry of the sequence, $+1$ in the same position as the third entry of the sequence, and so on alternately choosing ± 1. All the other entries of B are 0. Notice that all the row and column sums of B are 0. Let $A_+ = A + a_{i'j'}B$ and $A_- = A - a_{i'j'}B$. Notice that both A_+ and A_- are nonnegative matrices (because of the minimality property of $a_{i'j'}$) whose row and column sums are $+1$ (because the row and column sums of B are 0), so A_+ and A_- are doubly stochastic. Since $A = \frac{1}{2}A_+ + \frac{1}{2}A_-$ and $A_+ \neq A$, we conclude that A is not an extreme point of the set of doubly stochastic matrices.

The argument just presented shows that a given matrix is an extreme point of the compact convex set of doubly stochastic matrices if and only if it is a permutation matrix. The theorem follows from the fact that every point in a compact convex set is a convex combination of extreme points. □

Since there are exactly $n!$ distinct permutation matrices in M_n, Birkhoff's theorem ensures that any doubly stochastic matrix can be expressed as a convex combination of at most $N = n!$ permutation matrices. A more refined analysis shows that not more than $N = n^2 - 2n + 2$ terms are needed.

Problems

1. Let $A \in M_n$ be a nonnegative nonzero matrix with a positive eigenvector $x = [x_i]$, and let $D = \text{diag}(x_1, \ldots, x_n)$. Show that $\rho = \rho(A) > 0$, that

$Ax = \rho x$ by (8.1.30), and that $ADe = \rho De$, where $e \in \mathbf{R}^n$ has all entries $+1$. Conclude that A is similar (via a diagonal similarity matrix with positive main diagonal entries) to a positive multiple [namely $\rho(A)$] of a stochastic matrix. This observation permits many questions about nonnegative matrices with a positive eigenvector to be reduced to questions about stochastic matrices.

2. Show that the sets of stochastic and doubly stochastic matrices in M_n are compact convex sets.

3. Show that the sets of stochastic and doubly stochastic matrices in M_n each constitute a semigroup under matrix multiplication; that is, if $A, B \in M_n$ are (doubly) stochastic, then AB is (doubly) stochastic.

4. Show that a nonnegative matrix $A \in M_n$ is stochastic if and only if $Ae = e$.

5. Show that a 2-by-2 doubly stochastic matrix is symmetric with equal diagonal entries.

6. Use the ideas employed in the proof of (8.7.1) to (a) give an alternate, direct proof that does not use results of Appendix B and (b) give an algorithm for decomposing a doubly stochastic matrix as a convex combination of permutations. *Hint:* If A is not a permutation, use the sequence of entries indicated in the proof to produce a permutation, a positive multiple of which may be subtracted from A to leave a nonnegative matrix with equal row and column sums and at least one fewer nonzero entries. Now, repeat the argument on this matrix and continue.

7. Show that the decomposition in (8.7.1) is not unique.

8. If a doubly stochastic matrix A is reducible, show that A is actually permutation-similar to a matrix of the form $\begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}$. What may be said about A_1 and A_2?

Further Reading. The idea of the proof of (8.7.1) is contained in B. Saunders and H. Schneider, "Applications of the Gordon–Stiemke Theorem in Combinatorial Matrix Theory," *SIAM Rev.* 21 (1979), 528–541, where related facts may be found. For a discussion of the result that every doubly stochastic matrix $A \in M_n$ is a convex combination of at most $n^2 - 2n + 2$ permutation matrices, see M. Marcus and R. Ree, "Diagonals of Doubly Stochastic Matrices," *Quart. J. Math Oxford*, Ser. 2, 10 (1959), 295–302.

APPENDIX A

Complex numbers

A *complex number* has the form

$$z = a + ib$$

in which a and b are real numbers and i is a formal symbol satisfying the relation $i^2 = -1$. The real number a is called the *real part* of z and is denoted $\operatorname{Re} z$; the real number b is called the *imaginary part* of z and is denoted $\operatorname{Im} z$. The *complex conjugate* $\bar{z}$ of the complex number $z = a + ib$ is $\bar{z} = a - ib$. If $z_1 = a_1 + ib_1$ and $z_2 = a_2 + ib_2$ are complex numbers, then the binary operations of addition and multiplication are defined in the following natural ways in terms of the corresponding operations for real numbers:

$$z_1 + z_2 = (a_1 + a_2) + i(b_1 + b_2), \qquad z_1 z_2 = a_1 a_2 - b_1 b_2 + i(a_1 b_2 + a_2 b_1)$$

Thus, addition is the result of adding real parts and adding imaginary parts, and multiplication is the result of algebraic expansion together with the relation $i^2 = -1$. The additive inverse of $z = a + ib$ is $-z = -a + i(-b)$, and, as long as $z \neq 0 = 0 + i0$, the multiplicative inverse of z is

$$\frac{1}{z} = \frac{a - ib}{a^2 + b^2} = \frac{a}{a^2 + b^2} + i\left(\frac{-b}{a^2 + b^2}\right)$$

Subtraction and division of complex numbers z_1 and z_2 are defined by

$$z_1 - z_2 = z_1 + (-z_2), \qquad \frac{z_1}{z_2} = z_1\left(\frac{1}{z_2}\right) = \frac{z_1 \bar{z}_2}{z_2 \bar{z}_2}$$

The set of all complex numbers is denoted by **C**; the operations of addition and multiplication are commutative, and **C** constitutes a field under

these operations, with the real number $0=0+i0$ as additive identity and the real number $1=1+i0$ as multiplicative identity. The real numbers **R** form a subfield of **C**; the *absolute value* (or *modulus*) of z, denoted $|z|$, is defined by $|z|=+(z\bar{z})^{1/2}$, which is always a nonnegative real number. The quotient z_1/z_2 is then $(1/|z_2|^2)z_1\bar{z}_2$, if $z_2 \neq 0$. The operations of multiplication and complex conjugation are easily verified to commute, $\overline{z_1 z_2}=\bar{z}_1\bar{z}_2$, and the complex conjugate of the complex conjugate is the original complex number again. Since $\operatorname{Re} z=(1/2)(z+\bar{z})$ and $\operatorname{Im} z=(1/2i)(z-\bar{z})$, the real numbers are just those $z \in \mathbf{C}$ such that $\operatorname{Im} z=0$, or equivalently, $z=\bar{z}$ $(=\operatorname{Re} z)$.

Geometrically, the complex numbers **C** may be thought of as a plane with origin at 0 and a "real axis" and "imaginary axis." Thus, $z=a+ib$ may be identified with the ordered pair (a, b). The *real axis* $\{z: \operatorname{Im} z=0\}$ is just the usual real line, and the *imaginary axis* $\{z: \operatorname{Re} z=0\}$ is just i times the real line or all "pure imaginary" numbers. The projection of $z \in \mathbf{C}$ onto the real axis (imaginary axis) is $\operatorname{Re} z$ ($i \operatorname{Im} z$). Complex conjugation is reflection across the real axis, and $|z|$ is the Euclidean distance of z from the origin in the complex plane. The open (closed) *right half-plane* of **C** is $\{z \in \mathbf{C}: \operatorname{Re} z>(\geq) 0\}$, and the open (closed) *upper half-plane* of **C** is $\{z \in \mathbf{C}: \operatorname{Im} z>(\geq) 0\}$. The *unit disc* of **C** is $\{z \in \mathbf{C}: |z| \leq 1\}$, and the disc about $a \in \mathbf{C}$ of radius r is $\{z \in \mathbf{C}: |z-a| \leq r\}$.

The last paragraph described the complex plane **C** in terms of *rectangular coordinates.* The complex plane may also be described usefully in terms of *polar coordinates,* in which the position of $z \in \mathbf{C}$ in the plane is described in terms of the radius r of the circle about the origin on which z lies and the angle θ, measured counterclockwise around from the real line, of a directed ray from the origin on which z lies. The polar coordinates of a are then (r, θ), and the notation $z=re^{i\theta}$ is used, in which $e^{i\theta} \equiv \cos\theta+i\sin\theta$. The angle θ is the *argument* of z, written $\theta=\arg z$; since $e^{i\theta}=e^{i(\theta \pm 2\pi)}$, $\arg z$ is only determined mod 2π. If $z=a+ib$ in rectangular coordinates and $z=re^{i\theta}$ in polar coordinates, the transformation from polar to rectangular coordinates is

$$a=r\cos\theta, \qquad b=r\sin\theta$$

and from rectangular to polar coordinates when $r \neq 0$ is

$$r=|z|=(a^2+b^2)^{1/2}, \qquad \theta=\arcsin\frac{b}{r}=\arg z$$

in which we generally take $0 \leq \theta < 2\pi$. Circular objects are often easily described in polar coordinates. The unit disc in **C**, for example, is $\{re^{i\theta}: 0 \leq r \leq 1,\ 0 \leq \theta < 2\pi\}$.

APPENDIX B

Convex sets and functions

Let V be a vector space over a field that contains the real numbers. A *convex combination* of a selection $v_1, \ldots, v_k \in V$ of elements of V is a linear combination whose coefficients are nonnegative and sum to 1:

$$\alpha_1 v_1+\cdots+\alpha_k v_k; \qquad \alpha_1, \ldots, \alpha_k \geq 0, \qquad \sum_{i=1}^{k} \alpha_i=1$$

A subset K of V is said to be *convex* if any convex combination of any selection of elements from K lies in K. Equivalently, K is convex if all convex combinations of *pairs* of points in K are again in K. Geometrically, this may be interpreted as saying that the line segment joining any two points of K must lie in K; that is, K has no "dents" or "holes." A convex set K for which $\alpha x \in K$ whenever $\alpha>0$ and $x \in K$ is called a *convex cone* (equivalently, positive linear combinations from K are in K). It is straightforward to verify that both the set sum and the intersection of two convex sets (respectively, convex cones) is again a convex set (respectively, convex cone).

Now let V be a real or complex vector space with a given norm, so one can speak of open, closed, and compact sets in V. An *extreme point* of a closed convex set K is a point $z \in K$ that may be written as a convex combination of points from K in only a trivial way; that is, $z=\alpha x+(1-\alpha)y$, $0<\alpha<1$, $x, y \in K$, implies $x=y=z$. A closed convex set may have a finite number of extreme points (e.g., a polyhedron), infinitely many extreme points (e.g., a closed disc), or no extreme points (e.g., the closed upper half-plane in $\mathbf{R}^2$). A compact convex set always has extreme points, however. The *convex hull* of a set S of points in V, denoted $\operatorname{Co}(S)$, is simply the set of all convex combinations of all selections of points from S, or, equivalently, the "smallest" convex set (intersection of all convex sets)

containing S. The Krein–Milman theorem says that a compact convex set is the closure of the convex hull of its extreme points. A compact convex set is said to be *finitely generated* if it has finitely many extreme points, the extreme points being called *generators* of the convex set.

Now suppose V is a real inner product space with inner product $\langle \cdot, \cdot \rangle$. The *separating hyperplane theorem* states that if $K_1, K_2 \subseteq V$ are two given nonempty nonintersecting convex sets with K_1 closed and K_2 compact, then there exists a hyperplane H in V such that K_1 lies in one of the closed half-spaces determined by H while K_2 lies in the other; that is, H separates K_1 and K_2. A *hyperplane* H in V is just a translation of the orthogonal complement of a one-dimensional subspace of V: $H = \{x \in V: \langle x-p, q \rangle = 0\}$ for given vectors $p, q \in V$, $q \neq 0$. The hyperplane H determines two open *half-spaces:* $H^+ = \{x \in V: \langle x-p, q \rangle > 0\}$, $H^- = \{x \in V: \langle x-p, q \rangle < 0\}$. The sets $H_0^+ = H^+ \cup H$ and $H_0^- = H^- \cup H$ are the closed half-spaces determined by H. Thus, separation means that $K_1 \subseteq H_0^+$ and $K_2 \subseteq H_0^-$ for some vectors p, q. There are various strengthenings of the separation conclusion depending upon additional assumptions about the two convex sets. For example, if the closures of K_1 and K_2 do not intersect, then the separation may be taken to be strict; that is, $K_1 \subseteq H^+$, $K_2 \subseteq H^-$. The closure of the convex hull of any bounded set $S \subset V$ can be obtained as the intersection of all closed half-spaces that contain S.

In the event that V is the vector space $\mathbf{C}^n$ with complex inner product $\langle \cdot, \cdot \rangle$, hyperplanes and half-spaces are defined similarly, *except* that $\mathbf{C}^n$ must be identified with $\mathbf{R}^{2n}$ and $\langle \cdot, \cdot \rangle$ must be replaced with the real inner product $\mathrm{Re}\langle \cdot, \cdot \rangle$ as follows. Identify $x+iy \in \mathbf{C}^n$ with $\left[\begin{smallmatrix} x \\ y \end{smallmatrix}\right] \in \mathbf{R}^{2n}$, and note that $\mathrm{Re}\langle x_1+iy_1, x_2+iy_2 \rangle = \langle x_1, x_2 \rangle + \langle y_1, y_2 \rangle$ by conjugate linearity of the complex inner product. Then $\langle x_1, x_2 \rangle + \langle y_1, y_2 \rangle$ is the (real) inner product of $\left[\begin{smallmatrix} x_1 \\ y_1 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} x_2 \\ y_2 \end{smallmatrix}\right]$, and hyperplanes and half-spaces defined in $\mathbf{R}^{2n}$ have the appropriate geometric interpretation in $\mathbf{C}^n$.

A real valued function f defined on a convex set $K \subseteq V$ is said to be *convex* if

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y) \tag{*}$$

for all $0 < \alpha < 1$ and all $x, y \in K$, $y \neq x$. If the above inequality is always strict, then f is called *strictly convex.* If the above inequality is reversed for all $0 < \alpha < 1$ and all $x, y \in K$, $y \neq x$, then f is called *concave* (or *strictly concave* if it is reversed and always strict). Equivalently, a concave (respectively strictly concave) function is just the negative of a convex (respectively strictly convex) function. Geometrically, the chord joining any two function values $f(x)$ and $f(y)$ lies above (respectively below) the graph of a convex (respectively concave) function. A linear function is both convex and concave. In the case $V = \mathbf{R}^n$, and K an open set, the Hessian

$$H(x) \equiv \left[\frac{\partial^2 f}{\partial x_i \, \partial x_j}(x) \right]$$

which is a symmetric matrix in $M_n(\mathbf{R})$, exists almost everywhere in K for a convex function f and is necessarily positive semidefinite for points in K at which it exists. It is positive definite in the strictly convex case. Conversely, a function whose Hessian is positive semidefinite (respectively positive definite) throughout a convex set is convex (respectively strictly convex). Similarly, negative definiteness corresponds to concavity.

Optimization of convex and concave functions has some pleasant properties. On a compact convex set the maximum (respectively, minimum) of a convex (respectively, concave) function is attained at an extreme point. On the other hand, on a convex set, the set of points at which the minimum (respectively maximum) of a convex (respectively concave) function is attained is convex and any local minimum (respectively maximum) is a global minimum (respectively maximum). For example, a strictly convex function attains a minimum at at most one point of a convex set, and a critical point is necessarily a minimum.

Convex combinations of real numbers obey some simple but frequently useful inequalities. If $x_1, \ldots, x_k$ are given real numbers, then

$$\min_{1 \leq i \leq k} x_i \leq \sum_{i=1}^{k} \alpha_i x_i \leq \max_{1 \leq i \leq k} x_i$$

for any convex combination $\alpha_1, \alpha_2, \ldots, \alpha_k \geq 0$ and $\alpha_1 + \cdots + \alpha_k = 1$.

Consideration of certain simple convex functions $f(\cdot)$ of one variable on an interval leads to various classical inequalities. One can use induction to show that the defining two-point inequality (*) on the interval implies an n-point inequality

$$f\left(\sum_{i=1}^{n} \alpha_i x_i \right) \leq \sum_{i=1}^{n} \alpha_i f(x_i), \qquad n = 2, 3, \ldots \tag{**}$$

whenever $\alpha_i \geq 0$, $\alpha_1 + \cdots + \alpha_n = 1$, and all x_i are in the interval.

Application of (**) to the strictly convex function $f(x) = -\log x$ over the interval $(0, \infty)$ leads to the *weighted arithmetic–geometric mean inequality*

$$\sum_{i=1}^{n} \alpha_i x_i \geq \prod_{i=1}^{n} x_i^{\alpha_i}, \qquad x_i \geq 0$$

which contains the *arithmetic–geometric mean inequality*

$$\frac{1}{n} \sum_{i=1}^{n} x_i \geq \left(\prod_{i=1}^{n} x_i \right)^{1/n}, \qquad x_i \geq 0$$

when all $\alpha_i = 1/n$. Equality holds if and only if all x_i are equal.

Application of (**) with $f(x)=x^p$, $p>1$, over the interval $(0,\infty)$ leads to *Hölder's inequality*

$$\sum_{i=1}^{n} x_i y_i \le \left(\sum_{i=1}^{n} x_i^p\right)^{1/p} \left(\sum_{i=1}^{n} y_i^q\right)^{1/q}$$

where $x_i, y_i > 0$, $p>1$, and $1/p+1/q=1$. Equality holds if and only if the vectors $[x_i^p]$ and $[y_i^q]$ are dependent. If we take $p=q=2$, we obtain a version of the *Cauchy-Schwarz inequality*

$$\sum_{i=1}^{n} x_i y_i \le \left(\sum_{i=1}^{n} x_i^2\right)^{1/2} \left(\sum_{i=1}^{n} y_i^2\right)^{1/2}$$

Equality holds if and only if the vectors $[x_i]$ and $[y_i]$ are dependent. From Hölder's inequality one can deduce *Minkowski's inequality*

$$\left[\sum_{i=1}^{n} (x_i+y_i)^p\right]^{1/p} \le \left(\sum_{i=1}^{n} x_i^p\right)^{1/p} + \left(\sum_{i=1}^{n} y_i^p\right)^{1/p}$$

where $x_i, y_i > 0$ and $p \ge 1$. Equality holds if and only if the vectors $[x_i]$ and $[y_i]$ are dependent.

Further Readings. For more information about convex sets and geometry see [Val]. For more about convex functions and inequalities see [Boa] and [BB].

APPENDIX C

The fundamental theorem of algebra

One historical motivation for introducing the complex numbers **C** was that polynomials with real coefficients may have nonreal complex zeroes. For example, the quadratic formula reveals that the equation $x^2-2x+2=0$ has roots (solutions) $\{1+i, 1-i\}$. All zeroes of any polynomial with real coefficients, however, are contained in **C**. In fact, if the field of possible coefficients is extended to **C**, all zeroes of all polynomials with complex coefficients are still contained in **C**. Thus, **C** is an example of an *algebraically closed field*; that is, there is no field **F** such that **C** is a subfield of **F**, and such that there is a polynomial with coefficients from **C** and with a zero in **F** that is not in **C**.

The *fundamental theorem of algebra* states that any polynomial $p(x)$, of degree at least 1, with complex coefficients has at least one zero z [i.e., z is a root of the equation $p(x)=0$] among the complex numbers. Using synthetic division, if z is zero of $p(x)$, then $x-z$ divides $p(x)$; that is, $p(x)=(x-z)q(x)$, in which $q(x)$ is a polynomial with complex coefficients, whose degree is 1 smaller than that of $p(x)$. The zeroes of $p(x)$ are then those of $q(x)$, together with z. The following is a consequence of the fundamental theorem of algebra.

Theorem. A polynomial of degree $n \ge 1$ with complex coefficients has, counting multiplicities, exactly n zeroes among the complex numbers.

The multiplicity of a root z of $p(x)=0$ is the largest integer k for which $(x-z)^k$ divides $p(x)$, that is, the "number" of times z occurs as a root of $p(x)=0$. If a root z has multiplicity 3, then it is counted 3 times toward the number n of roots of $p(x)=0$. It follows that a polynomial with

complex coefficients may always be factored into a product of linear factors over the complex numbers.

If a polynomial $p(x)$ with *real* coefficients has some nonreal complex zeroes, however, they must occur in *conjugate pairs,* since, if $0=p(z)$, then $0=\bar{0}=\overline{p(z)}=p(\bar{z})$. It follows, since

$$(x-z)(x-\bar{z})=x^2-2\,\mathrm{Re}(z)x+|z|^2$$

that any real polynomial may be factored into a product of powers of linear and quadratic factors over the reals, each irreducible quadratic factor corresponding to a conjugate pair of complex roots.

Further Readings. For an elementary proof of the fundamental theorem of algebra, see [Chi].

APPENDIX D

Continuous dependence of the zeroes of a polynomial on its coefficients

It is an important fact, most readily proved using complex analysis, that the n zeroes of a polynomial of degree $n \geq 1$ with complex coefficients depend continuously upon the coefficients.

For $x \in \mathbf{C}^n$, let $f(x)=[f_1(x),\ldots,f_m(x)]^T$, in which $f_i:\mathbf{C}^n \to \mathbf{C}$, $i=1,\ldots,m$. The function $f:\mathbf{C}^n \to \mathbf{C}^m$ is *continuous* at x if each f_i is continuous at x, $i=1,\ldots,m$. The function $f_i:\mathbf{C}^n \to \mathbf{C}$ is continuous at x if, for each $\epsilon>0$ there is a $\delta>0$ such that if $\|y-x\|<\delta$, then $|f_i(y)-f_i(x)|<\epsilon$, where $\|\bullet\|$ is a vector norm on $\mathbf{C}^n$.

The continuous dependence result could be stated intuitively by saying that the function $f:\mathbf{C}^n \to \mathbf{C}^n$, which takes the n coefficients (all but the leading 1) of a monic polynomial of degree n to the n zeroes of the polynomial, is continuous. There is a problem, however; there is no simple way to define this function, since there is no natural way to define an ordering among the n zeroes. As a precise statement of the continuous dependence of the zeroes on the coefficients of a polynomial, we offer the following.

Theorem. Let $n \geq 1$ and let

$$p(x)=a_nx^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0, \qquad a_n \neq 0$$

be a polynomial with complex coefficients. Then, for every $\epsilon>0$, there is a $\delta>0$ such that for any polynomial

$$q(x)=b_nx^n+b_{n-1}x^{n-1}+\cdots+b_1x+b_0$$

satisfying $b_n \neq 0$ and

$$\max_{0 \le i \le n} |a_i - b_i| < \delta$$

we have

$$\min_{\tau} \max_{1 \le j \le n} |\lambda_j - \mu_{\tau(i)}| < \epsilon$$

where $\lambda_1, \ldots, \lambda_n$ are the zeroes of $p(x)$ and $\mu_1, \ldots, \mu_n$ are the zeroes of $q(x)$ in some order, counting multiplicities, and the minimum is taken over all permutations τ of $1, 2, \ldots, n$.

Thus, sufficiently small changes in the coefficients of a polynomial can lead only to small changes in any zero. This principle is of fundamental importance in matrix analysis because the coefficients of the characteristic polynomial $p_A(t)$ of a matrix $A \in M_n$ are continuous (in fact, polynomial) functions of the entries of A (1.2.11) and the zeroes of $p_A(t)$ are the eigenvalues of A. Since the composition of continuous functions is continuous, sufficiently small changes in the entries of A will cause only small changes in the coefficients of $p_A(t)$, which result in small changes in the eigenvalues. Thus, the eigenvalues of a square real or complex matrix depend continuously upon its entries.

Further Readings. For explicit bounds on the deviation ϵ between the zeroes of $p(x)$ and $q(x)$ in terms of the coefficient separation δ and the sizes of the coefficients, see L. Elsner, "On the Variation of the Spectra of Matrices," *Linear Algebra Appl.* 47 (1982), 127–138.

APPENDIX E

Weierstrass's theorem

Let V be a finite-dimensional real or complex vector space with norm $\|\cdot\|$. The *ball of radius* ϵ about $x \in V$ is $B_\epsilon(x) = \{y \in V\colon \|y - x\| \le \epsilon\}$. A subset $S \subseteq V$ is said to be *open* if for every $x \in S$ there is an $\epsilon > 0$ such that $B_\epsilon(x) \subseteq S$. A subset $T \subseteq V$ is said to be *closed* if the complement of T in V is open. A subset $S \subseteq V$ is called *bounded* if there is an $r > 0$ such that $S \subseteq B_r(0)$. Equivalently, T is closed if and only if the limit of any convergent (with respect to $\|\cdot\|$) sequence from T is in T, and S is bounded if S is contained in any ball of finite radius. A subset $S \subseteq V$ is *compact* if it is both closed and bounded.

For $S \subseteq V$, a function $f\colon S \to \mathbf{R}$ may or may not attain a (global) maximum or minimum value on S. However, under certain frequently occurring circumstances, we may be sure that f attains a maximum on S.

Theorem (Weierstrass). Let S be a compact subset of a finite-dimensional real or complex vector space V. If $f\colon S \to \mathbf{R}$ is a continuous function, then there exists a point $x_{\min} \in S$ such that

$$f(x_{\min}) \le f(x) \quad \text{for all} \quad x \in S$$

and there exists a point $x_{\max} \in S$ such that

$$f(x) \le f(x_{\max}) \quad \text{for all} \quad x \in S$$

That is, f attains its minimum and maximum on S. Of course, the values $\max_{x \in S} f(x)$ and $\min_{x \in S} f(x)$ may each be attained at more than one point of S. If either of the key assumptions (compact S and continuous f) of Weierstrass's theorem do not hold, the conclusion may fail. The

assumption that S is a subset of a finite-dimensional real or complex vector space is not essential, however. With a suitable definition of *compact*, Weierstrass's theorem holds for a continuous real-valued function on a compact subset of a general topological space.

References

Ait	A. C. Aitken. *Determinants and Matrices.* 9th ed. Oliver and Boyd, Edinburgh, 1956.
Bar 75	S. Barnett. *Introduction to Mathematical Control Theory*. Clarendon Press, Oxford, 1975.
Bar 79	S. Barnett. *Matrix Methods for Engineers and Scientists.* McGraw-Hill, London, 1979.
Bar 83	S. Barnett. *Polynomials and Linear Control Systems.* Dekker, New York, 1983.
BB	E. F. Beckenbach and R. Bellman. *Inequalities.* Springer-Verlag, New York, 1965
Bel	R. Bellman. *Introduction to Matrix Analysis.* 2d ed. McGraw-Hill, New York, 1970.
Boa	R. P. Boas, Jr. *A Primer of Real Functions.* 2d ed. Carus Mathematical Monographs, No. 13. Mathematical Association of America, Washington, D.C., 1972.
BPl	A. Berman and R. Plemmons. *Nonnegative Matrices in the Mathematical Sciences.* Academic Press, New York, 1979.
BSt	S. Barnett and C. Storey. *Matrix Methods in Stability Theory.* Barnes & Noble, New York, 1970.
CaLe	J. A. Carpenter and R. A. Lewis. *KWIC Index for Numerical Algebra.* U.S. Dept. of Commerce, Springfield, Va. Microfiche and printed versions available from National Technical Information Service, U.S. Dept. of Commerce, 5285 Port Royal Road, Springfield, VA 22161.
Chi	L. Childs. *A Concrete Introduction to Higher Algebra.* Springer-Verlag, Berlin, 1979.
Cul	C. G. Cullen. *Matrices and Linear Transformations.* Addison-Wesley, Reading, Mass., 1966.
Don	W. F. Donoghue, Jr. *Monotone Matrix Functions and Analytic Continuation.* Springer-Verlag, Berlin, 1974.

Fad V. N. Faddeeva. Trans. C. D. Benster. *Computational Methods of Linear Algebra.* Dover, New York, 1959.

Fan Ky Fan. *Convex Sets and Their Applications.* Lecture Notes, Applied Mathematics Division, Argonne National Laboratory, Summer 1959.

Fie M. Fieldler. *Spectral Properties of Some Classes of Matrices.* Lecture Notes, Report No. 75.01R. Chalmers University of Technology and the University of Göteborg, 1975.

Fra J. Franklin. *Matrix Theory.* Prentice-Hall, Englewood Cliffs, N.J., 1968.

Gan F. R. Gantmacher. *The Theory of Matrices.* 2 vols. Chelsea, New York, 1959.

Gant F. R. Gantmacher. *Applications of the Theory of Matrices.* Interscience, New York, 1959.

GKr F. R. Gantmacher and M. G. Krein. *Oszillationsmatrizen, Oszillationskerne, und kleine Schwingungen mechanische Systeme.* Akademie-Verlag, Berlin, 1960.

GLR 82 I. Gohberg, P. Lancaster, and L. Rodman. *Matrix Polynomials.* Academic Press, New York, 1982.

GLR 83 I. Gohberg, P. Lancaster, and L. Rodman. *Matrices and Indefinite Scalar Products.* Birkhäuser-Verlag, Boston, 1983.

Grah A. Graham. *Kronecker Products and Matrix Calculus with Applications.* Horwood, Chichester, U.K., 1981.

Gray F. A. Graybill. *Matrices with Applications to Statistics.* 2d ed. Wadsworth, Belmont, Calif., 1983.

Gre W. H. Greub. *Multilinear Algebra.* 2d ed. Springer-Verlag, New York, 1978.

GVl G. Golub and C. VanLoan. *Matrix Computations.* Johns Hopkins University Press, Baltimore, 1983.

Hal 58 P. R. Halmos. *Finite-Dimensional Vector Spaces.* Van Nostrand, Princeton, N.J., 1958.

Hal 67 P. R. Halmos. *A Hilbert Space Problem Book.* Van Nostrand, Princeton, N.J., 1967.

HJ R. Horn and C. Johnson. *Topics in Matrix Analysis.* Cambridge University Press, Cambridge, 1989.

HKu K. Hoffman and R. Kunze. *Linear Algebra.* 2d ed. Prentice-Hall, Englewood Cliffs, N.J., 1971.

Hou 64 A. S. Householder. *The Theory of Matrices in Numerical Analysis.* Blaisdell, New York, 1964.

Hou 72 A. S. Householder. *Lectures on Numerical Algebra.* Mathematical Association of America, Buffalo, N.Y., 1972.

HSm M. W. Hirsch and S. Smale. *Differential Equations, Dynamical Systems, and Linear Algebra.* Academic Press, New York, 1974.

Jac N. Jacobson. *The Theory of Rings.* American Mathematical Society, New York, 1943.

Kap I. Kaplansky. *Linear Algebra and Geometry: A Second Course.* Allyn & Bacon, Boston, 1969.

Kar S. Karlin. *Total Positivity.* Stanford University Press, Stanford, Calif., 1960.

Kel R. B. Kellogg. *Topics in Matrix Theory.* Lecture Notes, Report No. 71.04, Chalmers Institute of Technology and the University of Göteberg, 1971.

Kow H. Kowalsky. *Lineare Algebra.* 4th ed. deGruyter, Berlin, 1969.

LaH C. Lawson and R. Hanson. *Solving Least Squares Problems.* Prentice-Hall, Englewood Cliffs, N.J., 1974.

Lan P. Lancaster. *Theory of Matrices.* Academic Press, New York, 1969.

LaTi P. Lancaster and M. Tismenetsky. *The Theory of Matrices With Applications.* 2d ed. Academic Press, New York, 1985.

Mac C. C. MacDuffee. *The Theory of Matrices.* Chelsea, New York, 1946.

Mar M. Marcus. *Finite Dimensional Multilinear Algebra.* 2 vols. Dekker, New York, 1973–75.

Mir L. Mirsky. *An Introduction to Linear Algebra.* Clarendon Press, Oxford, 1963.

MMi M. Marcus and H. Minc. *A Survey of Matrix Theory and Matrix Inequalities.* Allyn & Bacon, Boston, 1964.

MOl A. W. Marshall and I. Olkin. *Inequalities: Theory of Majorization and Its Applications.* Academic Press, New York, 1979.

Mui T. Muir. *The Theory of Determinants in the Historical Order of Development.* 4 vols. MacMillan, London, 1906, 1911, 1920, 1923; Dover, New York, 1966. *Contributions to the History of Determinants, 1900–1920.* Blackie, London, 1930.

Ner E. Nering. *Linear Algebra and Matrix Theory.* 2d ed. Wiley, New York, 1963.

New M. Newman. *Integral Matrices.* Academic Press, New York, 1972.

Nob B. Noble. *Applied Linear Algebra.* Prentice-Hall, Englewood Cliffs, N.J., 1969.

Per S. Perlis. *Theory of Matrices.* Addison-Wesley, Reading, Mass., 1952.

Rog G. S. Rogers. *Matrix Derivatives.* Lecture Notes in Statistics, Vol. 2. Dekker, New York, 1980.

Rud W. Rudin. *Principles of Mathematical Analysis.* 3rd ed. McGraw-Hill, New York, 1976.

Sen E. Seneta. *Nonnegative Matrices.* Wiley, New York, 1973.

Ste G. W. Stewart. *Introduction to Matrix Computations.* Academic Press, New York, 1973.

Str G. Strang. *Linear Algebra and Its Applications.* Academic Press, New York, 1976.

STy D. A. Suprenenko and R. I. Tyshkevich. *Commutative Matrices.* Academic Press, New York, 1968.

Tod J. Todd (ed.). *Survey of Numerical Analysis.* McGraw-Hill, New York, 1962.

TuA H. W. Turnbull and A. C. Aitken. *An Introduction to the Theory of Canonical Matrices.* Blackie, London, 1932.

Tur H. W. Turnbull. *The Theory of Determinants, Matrices and Invariants.* Blackie, London, 1950.

Val F. A. Valentine. *Convex Sets.* McGraw-Hill, New York, 1964.

Var R. S. Varga. *Matrix Iterative Analysis.* Prentice-Hall, Englewood Cliffs, N.J., 1962.

Wed J. H. M. Wedderburn. *Lectures on Matrices.* American Mathematical Society Colloquium Publications XVII. American Mathematical Society, New York, 1934.

Wie H. Wielandt. *Topics in the Analytic Theory of Matrices.* Lecture Notes prepared by R. Meyer. Department of Mathematics, University of Wisconsin, Madison, 1967.

Wil J. H. Wilkinson. *The Algebraic Eigenvalue Problem.* Clarendon Press, Oxford, 1965.

Notation

$\mathbf{R}$	the real numbers
$\mathbf{R}^n$	real vector space of real n-vectors, $M_{n,1}(\mathbf{R})$
$\mathbf{C}$	the complex numbers
$\mathbf{C}^n$	complex vector space of complex n-vectors, $M_{n,1}(\mathbf{C})$
$\mathbf{F}$	a field (usually $\mathbf{R}$ or $\mathbf{C}$)
$\mathbf{F}^n$	vector space (over $\mathbf{F}$) of n-vectors with entries from $\mathbf{F}$, $M_{n,1}(\mathbf{F})$
$M_{m,n}(\mathbf{F})$	m-by-n matrices with entries from $\mathbf{F}$
$M_{m,n}$	m-by-n complex matrices, $M_{m,n}(\mathbf{C})$
M_n	n-by-n complex matrices, $M_{n,n}(\mathbf{C})$
A, B, C, etc.	matrices; $A = [a_{ij}] \in M_{m,n}(\mathbf{F})$
x, y, z, etc.	column vectors; $x = [x_i] \in \mathbf{F}^n$
I	identity matrix in $M_n(\mathbf{F})$
0	zero scalar, vector, or matrix
$\bar{A}$	matrix of complex conjugates of entries of $A \in M_{m,n}(\mathbf{C})$
A^T	transpose of $A \in M_{m,n}(\mathbf{F})$
A^*	Hermitian adjoint of $A \in M_{m,n}(\mathbf{C})$, $\bar{A}^T$
A^{-1}	inverse of a nonsingular $A \in M_n(\mathbf{F})$
$A^{1/2}$	unique positive semidefinite square root of a positive semidefinite $A \in M_n$
$\|A\|$	matrix of absolute values of entries of $A \in M_{m,n}$
$A^\dagger$	Moore–Penrose generalized inverse of $A \in M_{m,n}$
$\operatorname{adj} A$	classical adjoint (adjugate) of $A \in M_n(\mathbf{F})$
$\mathcal{B}$	a basis of a vector space
e_i	ith standard basis vector in $\mathbf{F}^n$ (usually)
$[v]_{\mathcal{B}}$	$\mathcal{B}$ coordinate representation of a vector v
${}_{\mathcal{B}_2}[T]_{\mathcal{B}_1}$	$\mathcal{B}_1$–$\mathcal{B}_2$ basis representation of a linear transformation T

$\binom{n}{k}$	binomial coefficient, $n!/[k!(n-k)!]$
$p_A(t)$	characteristic polynomial of $A \in M_n(\mathbf{F})$
$\kappa(A)$	condition number (for inversion, with respect to a given matrix norm) of a nonsingular $A \in M_n$
det A	determinant of $A \in M_n(\mathbf{F})$
$\oplus$	direct sum
$\Gamma(A)$	directed graph of $A \in M_n(\mathbf{F})$
$\|\cdot\|^D$	dual norm of a vector norm $\|\cdot\|$
$f^D(\cdot)$	dual norm of a pre-norm $f(\cdot)$
λ	eigenvalue of $A \in M_n$ (usually)
$\{\lambda_i(A)\}$	set of eigenvalues (spectrum) of $A \in M_n$; if A is Hermitian, one usually takes $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$
$n!$	factorial, $n(n-1)(n-2)\ldots 2\cdot 1$
$G(A)$	Geršgorin region of $A \in M_n$
$GL(n,F)$	group of nonsingular matrices in $M_n(\mathbf{F})$
$A \circ B$	Hadamard product of $A, B \in M_{m,n}(\mathbf{F})$
$\gamma(A)$	index of primitivity of a primitive $A \in M_n$
$M(A)$	indicator matrix of $A \in M_{m,n}(\mathbf{F})$
$J_k(\lambda)$	Jordan block of size k with eigenvalue λ
$\otimes$	Kronecker (tensor) product
$q_A(t)$	minimal polynomial of $A \in M_n(\mathbf{F})$
$\|\cdot\|_1$	l_1 (sum) norm of $\mathbf{C}^n$; l_1 matrix norm on M_n
$\|\cdot\|_2$	l_2 (Euclidean) norm on $\mathbf{C}^n$; l_2 (Frobenius) matrix norm on M_n
$\|\cdot\|_\infty$	l_∞ (max) norm on $\mathbf{C}^n$; l_∞ vector norm on M_n
$\|\cdot\|_p$	l_p norm on $\mathbf{C}^n$
$\|\|\cdot\|\|_1$	maximum column sum matrix norm on M_n
$\|\|\cdot\|\|_2$	spectral matrix norm on M_n
$\|\|\cdot\|\|_\infty$	maximum row sum matrix norm on M_n
$r(A)$	numerical radius of $A \in M_n$ (usually)
$\perp$	orthogonal complement
per A	permanent of $A \in M_n(\mathbf{F})$
rank A	rank of $A \in M_{m,n}(\mathbf{F})$
sgn	signum of a permutation
$\{\sigma_i(A)\}$	set of singular values of $A \in M_{m,n}$; one usually takes $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min\{m,n\}} \geq 0$
$\sigma_1(A)$	largest singular value of $A \in M_{m,n}$, $\|\|A\|\|_2$
Span S	span of a subset S of a vector space
$\rho(A)$	spectral radius of $A \in M_n$
$\sigma(A)$	spectrum (set of eigenvalues) of $A \in M_n$
$A(\alpha,\beta)$	submatrix of $A \in M_{m,n}(\mathbf{F})$ determined by the index sets α, β
tr A	trace of $A \in M_n(\mathbf{F})$

Index